PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68, C12N 15/10

A1

(11) International Publication Number: **WO 98/31838**

(43) International Publication Date: 23 July 1998 (23.07.98)

(21) International Application Number: PCT/US98/00965

(22) International Filing Date: 15 January 1998 (15.01.98)

(30) Priority Data:
08/784,208 15 January 1997 (15.01.97) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 08/784,208 (CIP)
Filed on 15 January 1997 (15.01.97)

(71) Applicant *(for all designated States except US)*: CHUGAI PHARMACEUTICAL CO., LTD. [JP/JP]; 1–9, Kyobashi 2–chome, Chuoku, Tokyo 104 (JP).

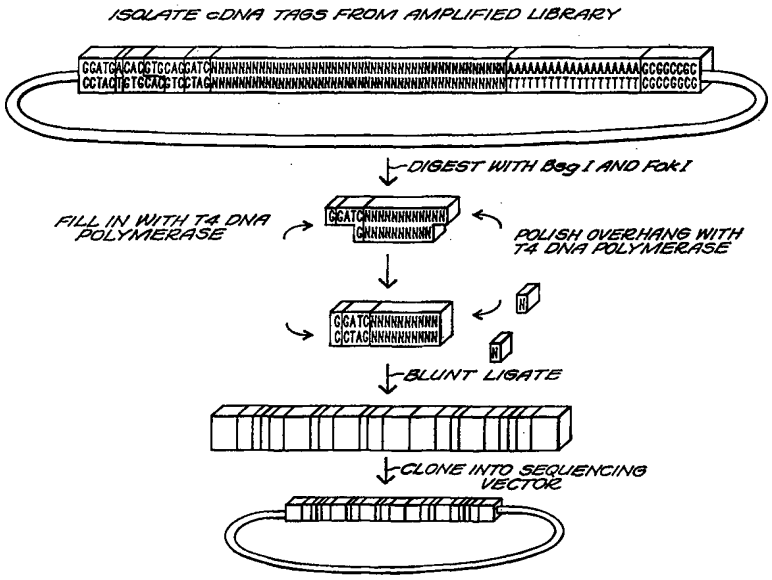
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: SPINELLA, Dominic, G. [US/US]; 7026 Via Calafia, La Costa, CA 92009 (US). SAJJADI, Fereydoun, G. [CA/US]; 1548 Plumtree Drive, Encinitas, CA 92024 (US).

(74) Agents: DECONTI, Giulio, A., Jr. et al.; Lahive & Cockfield, LLP, 28 State Street, Boston, MA 02109 (US).

(81) Designated States: AU, CA, JP, KR, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD FOR ANALYZING QUANTITATIVE EXPRESSION OF GENES



(57) Abstract

The present invention provides novel methods for identifying gene expression patterns in mRNA populations. The methods are useful for determining differential gene expression among various cells or tissues, including cells or tissues of a target organism. The invention also provides methods of determining the frequency of gene expression in mRNA populations, thus providing a method of comparing gene expression frequency among various cells or tissues. The present invention also provides methods for isolating genes corresponding to tag sequences identified according to the methods of the present invention. Furthermore, sequences that are identified according to the present invention may be used to diagnose the presence of disease.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD FOR ANALYZING QUANTITATIVE EXPRESSION OF GENES

Field of the Invention

The present invention relates to novel methods for identifying gene expression patterns in cells and tissues, methods for determining the frequency of gene expression in cells and tissues, including cells or tissues of a target organism, and vectors used for identifying gene expression patterns. Target organisms include humans, animals and plants. The present invention also provides methods for isolating genes corresponding to tag sequences identified according to the methods of the present invention. The present invention also relates to methods for diagnosing diseases related to differential gene expression and to methods for determining the effects of drugs on gene expression.

Background of the Invention

The human genome contains approximately 100,000 genes, however, in any given cell, only a fraction of these genes are expressed. Thus, in each cell type, only a fraction of human genes are expressed at any one time. Each gene is expressed at a precise time and at a precise level.

Automated DNA sequencers have made it easier to determine the sequence of the genome of an organism; the genomic sequences of *Haemophilus influenzae*, *Mycoplasma genitalium*, and *Caenorhabditis elegans* have been published leading to the possibility that the genomic sequence of other higher organisms, such as humans, may be obtained (Fleischmann, R.D. et al. (1995) *Science* 269:496; Fraser, C.M. et al. (1995) *Science* 270:397; Hodgkin, J. et al. (1995) *Science* 270:410). However, the information derived from this technology still does not answer the question of which of these genes are expressed at any one time in any given cell. This information is crucial to determine how cells are differentiated from each other, how cells age, and the causes and effects of many diseases.

A typical mammalian cell of a given lineage expresses approximately 20,000-30,000 of the 100,000 odd germ line genes carried in its genome. Almost all cells universally express many of the same genes, which are called "housekeeping" genes. Examples of housekeeping genes include genes encoding enzymes involved in glycolysis or proteins involved in cell structure. However, it is the non-universally expressed genes that differentiate cells from each other. As cells mature into differentiated cells, certain non-constitutively expressed genes are turned on and off at different stages. Thus, the differences in gene expression patterns between cells make, for example, a nerve cell different from a blood cell.

Furthermore, the intracellular concentration of a non-constitutively expressed gene product can be modulated by the induction or repression of gene expression in response to environmental signals. Thus, the relative concentration of gene products within a given cell type can be indicative of the state of the cell.

Even within a single cell, the level of expression can vary a great deal from one gene to the next. In a typical cell, there are perhaps 200,000 mRNA molecules which represent 20,000-30,000 different transcribed sequences, present in the cytoplasm. A few of these transcript sequences may be present in high abundance, with thousands of copies or more present per cell. For example, up to 70% of the total mRNA in an antibody secreting plasma cell is represented by immunoglobulin mRNA. Other genes, typically housekeeping genes such as actin or glucose-6-phosphate dehydrogenase, are present at medium abundance with approximately 100-1,000 copies per cell. However, more than 90% of gene transcripts, are present in low abundance at a level of less than 10-15 copies per cell.

Under abnormal cellular conditions such as those in individuals with diseases or disorders, the pattern of gene expression within individual cells may be changed compared to the expression pattern seen under normal non-disease conditions. A change in gene expression may be an effect or the cause of a disease or other abnormality, such as in, for example, a tumor cell. Whereas some diseases may be understood as caused by mutations in particular genes and thus could potentially be detected by examining the genomic sequence, many diseases and disorders involve a malfunction in the level of expression of genes which cannot be detected by sequencing the genome but can only be detected by identifying the gene expression patterns of the cells. Therefore, in order to understand the function of specific cell types in an organism or to understand the progression of disease, it is necessary to understand the expression status of individual genes within these specific cell types at different stages of the organism's development.

One way researchers have attempted to answer these questions is to isolate proteins from various cells and to compare the abundance of each of these proteins. In one approach, proteins are purified from the cells and their abundance is compared. However, this approach is limited by difficulties in devising equally efficient methods of purifying different proteins. This approach is also limited to known proteins. In another approach, two-dimensional gel electrophoresis is used to compare protein expression, but this may lead to difficulties in resolving all of the proteins in the cell and in detecting proteins that are produced at a very low level (*See* Kahn, P. (1995) *Science* 270:369).

Other methods of determining peptide expression in an mRNA population involve the use of antibodies to probe populations of peptides produced from mRNA pools. Thus, "libraries" of synthetic polypeptides corresponding to the polypeptides

coded for by mRNA molecules are produced and then probed by individual antibodies. This method does not provide for a detection of all of the polypeptides produced by the mRNA at one time as it may not detect low levels of expression. Moreover, the method is limited to available antibodies. This method is described in, for example, U.S. Patent No. 5,242,798, issued September 7, 1993, and in U.S. Patent 4,900,811, issued February 13, 1990.

Furthermore, in all of these protein detection methods, once a particular protein difference has been determined, the protein must still be partially sequenced and cloned in order to determine the gene that is responsible for expression of the protein. Alternatively, the protein must be sequenced and compared to a "proteome" database (Kahn, P. (1995) *Science* 270:369). Moreover, determining gene expression patterns by looking at purified proteins from the cell is a method of looking at secondary and tertiary effects of gene expression -- translation of mRNA into protein, and post-translational modification -- and not the primary effect -- transcription of DNA sequences into mRNA. Detecting protein expression levels, furthermore, does not take into account the possibility that proteins may be degraded after translation and that the difference in protein expression is not actually due to a difference in gene expression.

Researchers have also focused on detecting changes in expression of individual mRNAs. One method involves subtractive hybridization, but this method does not have sufficient resolution to detect RNAs that are expressed at very low levels. Lee, S.W. et al. (1991) *Proc. Natl. Acad. Sci.* USA 88:2825. Another method involves a microarray hybridization assay where cDNA is prepared from two mRNA populations, labeled with two different colors, and used to hybridize to microscope slides to which a cDNA library has been fixed; differential hybridization is then identified by determining whether the sample fluoresces (*See*, Nowak, R. (1995) *Science* 270:368; Schena et al. (1995) *Science* 270:467). Recently, researchers have focused on short specific sequences of each mRNA called "tags" which are specific for a particular mRNA in the cell and are sufficient to identify the expression of a particular mRNA. These tags are analogous to sequences found at sequence tag sites (STS) that have been used to identify and map genomic markers (Olson et al. (1989) *Science* 245:1434). In one such method, randomly chosen cDNA clones are made from mRNAs of a particular tissue. This bulk method of producing cDNAs results in a database of "expressed sequence tags" (Adams, M.D. et al. (1991) *Science*, 255:1651; Adams, M.D. (1992) *Nature* 355:632-634).

Other methods have focused on using the polymerase chain reaction (PCR) to define tags and to attempt to detect differentially expressed genes. Many groups have used the PCR method to establish databases of mRNA sequence tags which could conceivably be used to compare gene expression among different tissues (Williams,

J.G.K. (1990) *Nucl. Acids Res.* 18:6531; Welsh, J., et al. (1990) *Nucl. Acids Res.* 18:7213; Woodward, S.R. (1992) *Mamm. Genome* 3:73; Nadeau, J.H. (1992) *Mamm. Genome* 3:55). This method has also been adapted to compare mRNA populations in a process called mRNA differential display. In this method, the results of PCR synthesis are subjected to gel electrophoresis, and the bands produced by two or more mRNA populations are compared. Bands present on an autoradiograph of one gel from one mRNA population, and not present on another, correspond to the presence of a particular mRNA in one population and not in the other, and thus indicate a gene that is likely to be differentially expressed. Messenger RNA derived from two different types of cells is compared by using arbitrary oligonucleotide sequences of ten nucleotides (random 10-mers) as a 5' primer and one of a set of 12 oligonucleotides complimentary to the poly A tail as a 3' "anchor primer." These primers are then used to amplify partial sequences of mRNAs with the addition of radioactive deoxyribonucleotides. These amplified sequences are then resolved on a sequencing gel such that each sequencing gel has a sequence of 50-100 mRNAs. The sequencing gels are then compared to each other to determine which amplified segments are expressed differentially (Liang, P. et al. (1992) *Science* 257:967; *See also* Welsh, J. et al. (1992) *Nucl. Acid Res.* 20:4965; Liang, P. et al. (1993) *Nucl. Acids Res.* 3269).

Another method based on using PCR to detect the expression of mRNAs relies on the use of 12 anchor primers which hybridize to the poly A tract and two restriction endonucleases, one that cleaves at a 4 nucleotide sequence within the cDNA sequence that corresponds to the mRNA, and another restriction endonuclease which recognizes a single site within each anchor primer. The cDNA derived from the mRNA in each of the 12 pools is then inserted into a vector, downstream from a promoter, and used to transform host cells in order to amplify the vector containing the cDNA insert. "cRNA" antisense transcripts are then made, driven by the promoter, which are then amplified using PCR. The PCR reaction is carried out with 16 or more different primers, in 16 different subpools. Thus, with 12 different anchor primers, 192 subpools are required per mRNA sample. The results of the PCR are then resolved on a sequencing gel (WO 95/13369, published May 18, 1995).

Another method for analyzing gene expression, referred to as Serial Analysis of Gene Expression (or SAGE), utilizes dimerized tags (termed "ditags") and concatenation of ditags for sequence analysis of expressed genes (Velculescu, V.E. et al. (1995) *Science* 270:484; US Patent No. 5,695,937). In this method, a cDNA copy of mRNA is made using a poly dT primer which is usually biotinylated. The cDNA copy is then made double-stranded and then cut with an "anchoring enzyme" which generally recognizes a four base pair sequence present in each cDNA. The biotinylated cDNA is

then bound to streptavidin beads to remove the rest of the sequence. This results in a cDNA copy of a portion of the 3' end of the messenger RNA linked to a streptavidin bead. The population of cDNAs linked to streptavidin beads is usually divided in half. Each half is then ligated to one of two oligonucleotide linkers containing a restriction endonuclease recognition site for a restriction endonuclease that cleaves DNA at a site different than the recognition site (e.g., a Type IIs restriction endonuclease), referred to as the "tagging enzyme", resulting in cleavage at a site within the cDNA copy of the mRNA sequence. The ends of the cDNA sequences are ligated together in pairs in a "tail to tail" manner to make a population of "ditags" that include an oligonucleotide linker at the 5' end and another oligonucleotide linker at 3' end. The ditags are generally amplified with PCR using primers specific to the linkers. The PCR- amplified regions are cleaved with the anchoring enzyme and concatenated together into a series of ditags punctuated by the sequence of the anchoring enzyme recognition site. The concatenated ditags can be sequenced directly or cloned into a vector and sequenced, and the sequences of the ditags are then compared to known sequences to identify expressed genes.

The use of PCR results in problems of reproducibility and requires the use of other complicated steps, including the preparation and annealing of PCR primers, to a method of detecting gene expression patterns. Moreover, these PCR-based methods do not necessarily detect differences in the frequency of gene expression.

The abundance of a PCR product after amplification is influenced by many factors in addition to starting template abundance. Sequence specific differences in "amplification efficiency" are well known to give rise to artifactual differences quantity of PCR product in the absence of real differences in starting template. Moreover, even repetitive amplification of the same template preparation has been reported to produce product yields that can vary by as much as 6-fold (Gilliand et al. in: PCR Protocols. Academic Press, pp 60-69 (1990)). Hence, any PCR-based method that attempts to infer starting template abundance from the quantity of product produced by amplification requires stringent co-amplification controls. In the above cited "SAGE" technique, all cDNA "tags" that happen to have a highly amplifiable sequence (e.g., an AT-rich sequence) will be over represented while those that have "difficult" sequences (e.g., GC-rich palindromic sequences) will be under-represented after the PCR step. The use of "ditags" fails to rectify all of the reliability problems involved in using SAGE to determine starting template abundance. Excluding any ditag that is repetitively isolated fails to eliminate all of the over-represented tag sequences. Artificially enhanced "amplifiability" may be the result of just one of the tags -- in which case any ditag

containing the individual member would be over-represented. Moreover, this exclusion does nothing about sequences which are artificially under-represented.

Thus, there is a need for a simple and reproducible method for detecting gene expression, identifying genes, and gene expression patterns in individual cells or tissues as well as a method for determining the frequency of gene expression in these cells or tissues.

Summary of the Invention

The present invention provides a method for tagging and identifying all of the expressed genes in a given cell population. This method thus allows even mRNAs with low copy number to be detected. By comparing gene expression profiles among cells, this method can be used to identify individual genes whose expression is associated with a pathological phenotype. Using high throughput DNA sequencing and associated information system support to analyze such DNA sequencing, the method of the present invention also permits the generation of global gene expression profiles in a reasonable length and time. Thus, the present invention provides a simple and rapid method of obtaining sufficient data to use in an information system known to those of skill in the art to obtain global gene expression profile and identify genes of interest.

The present invention employs methods for identifying gene expression patterns in an mRNA population. A preferred use of the methods of the present invention is to identify differential gene expression patterns among two or more cells or tissues. Thus, using the methods of the present invention, one can identify a gene or genes that is (are) expressed in any given cell type, tissue, or target organism at a different level from that in another cell type, tissue, or target organism. The methods of the present invention can also be used to identify differential gene expression at different stages of development in the same cell-type or tissue-type, and to identify changes in gene expression patterns in diseased or abnormal cells. Furthermore, the invention can be used to detect changes in gene expression patterns due to changes in environmental conditions or to treatment with drugs. Three different embodiments of these methods are described below.

In one aspect of the invention there is provided a method for identifying gene expression patterns in an mRNA population. The method includes preparing double-stranded cDNAs from an mRNA population using a primer, e.g., an oligo dT sequence linked at the 5' end of the oligo dT sequence to a cleavage site for a "priming" restriction endonuclease, and cleaving the double-stranded cDNAs with a first restriction endonuclease, which cleaves at a site within the cDNA sequence and not within the primer, to obtain cDNA inserts. The cDNA inserts are inserted into the insertion sites of cloning vectors to obtain DNA constructs, wherein each cloning vector includes a

second restriction endonuclease recognition sequence 5' to the insertion site such that digestion of the DNA construct with the second restriction endonuclease cleaves the DNA construct at a site within the cDNA insert, and a third restriction endonuclease recognition sequence 5' to or overlapping with the second restriction endonuclease recognition sequence. DNA constructs are amplified, e.g., in a suitable host cell, e.g., *E. coli*, and isolated. After isolation, the amplified DNA constructs are digested with the second and the third restriction endonuclease to obtain tags. The nucleotide sequence of the tags is then obtained to identify gene expression patterns in the mRNA population.

In preferred aspects, the nucleotide sequence of the tags is obtained by ligating the tags to obtain ligated tag arrays of at least about 10 tags, more preferably of at least about 40 tags, inserting the ligated tag arrays into a sequencing vector, and sequencing the ligated tag arrays. In one embodiment, the first restriction endonuclease recognizes a sequence of four bases; the second restriction endonuclease is a Type IIs restriction endonuclease; and the third restriction endonuclease recognition sequence is located about 10 to 40 nucleotides 5' of the second restriction endonuclease cleavage site. In another embodiment, the first restriction endonuclease recognizes a sequence of four bases; the second restriction endonuclease is a Type IIs restriction endonuclease; and the third restriction endonuclease recognition sequence overlaps the second restriction endonuclease recognition sequence. In one embodiment the population of double stranded cDNA is prepared by digestion of the double-stranded cDNA with a priming restriction endonuclease to obtain cDNA inserts comprising the priming restriction endonuclease cleavage sequence introduced at a 3' end of the double-stranded cDNA when the cDNA is digested with the priming restriction endonuclease. The priming restriction endonuclease can recognize sequences consisting of more than six bases, preferably it recognizes an eight-base palindromic sequence. Most preferably, the priming restriction endonuclease is NotI. It is also preferable that the first restriction endonuclease have a high probability of recognizing a sequence within each cDNA. Thus, in preferred aspects of the invention, the first restriction endonuclease recognizes a sequence consisting of less than six bases. More preferably, the first restriction endonuclease recognizes a sequence consisting of four bases. A preferred restriction endonuclease is MboI. It is also preferred that the second restriction endonuclease cleaves DNA at a site downstream of the recognition site for the endonuclease such that digestion of the vector with the second restriction endonuclease results in cleavage of the cDNA insert at a site within the sequences corresponding to the copied mRNA. Preferably, the second restriction endonuclease is a Type IIs restriction endonuclease. More preferably, the second restriction endonuclease cleaves DNA 10-14 bases 3' to the

recognition sequence. More preferably the second restriction endonuclease is a Type IIS restriction endonuclease. Most preferably, the second restriction endonuclease is BsgI. In other preferred aspects, the third restriction endonuclease recognition sequence is within about 20 to 40, more preferably about 10 to 15, nucleotides 5' of the second restriction endonuclease cleavage sequence. A cleavage site at a relatively short distance from the second restriction endonuclease cleavage sequence is preferable in order to maximize the number of tags that may be inserted into a sequencing vector. Preferably, the third restriction endonuclease recognition sequence is within about 10 to 15 nucleotides 5' of the third restriction endonuclease cleavage site. In one embodiment, the recognition sequence of the third restriction endonuclease overlaps with the recognition sequence of the second restriction endonuclease. Preferably, the third restriction endonuclease recognition sequence is within the second restriction endonuclease recognition sequence. It is also preferable that cleavage of the DNA with the third restriction endonuclease leaves a blunt end. Preferably, the second restriction endonuclease is BsgI and the third restriction endonuclease is PmlI. In a more preferred embodiment, the third restriction site is a Type IIs site in which the cleavage site is located immediately 5' to the second restriction cleavage site. Most preferably, the third restriction site is FokI.

In a preferred aspect, a method is provided for identifying gene expression patterns in an mRNA population. The method includes preparing double-stranded cDNAs from a mRNA population using a primer, wherein the primer comprises an oligo dT sequence linked at the 5' end of the oligo dT sequence to a NotI cleavage site and cleaving the double-stranded cDNAs with NotI and with MboI to obtain cDNA inserts. The cDNA fragments are inserted into an insertion site of a cloning vector to obtain DNA constructs, wherein the cloning vector further comprises: (i) a BsgI recognition sequence 5' to the insertion site such that digestion of the DNA construct with BsgI cleaves the DNA construct at a site within the cDNA insert, and (ii) a FokI recognition sequence which is located 5' to the BsgI recognition sequence. The DNA constructs containing the cDNA inserts are amplified in a suitable host and isolated. After isolation, the amplified DNA constructs are digested with BsgI and FokI to obtain tags. The tags are treated with T4 DNA polymerase to obtain blunt ends and then ligated using DNA ligase to obtain ligated tag arrays of at least about 30-60 tags. The ligated tag arrays are inserted into a sequencing vector and sequenced. The sequences of individual tags within the ligated tag arrays are compared to known gene sequences to identify gene expression patterns in the mRNA population.

In preferred aspects, the tags have blunt 5' and 3' ends. Preferably, the tags are treated with a DNA polymerase after restriction enzyme digestion, such as, for example,

T4 DNA polymerase, to obtain tags having blunt 5' and 3' ends. To aid in sequencing the tags, tags are preferably ligated together using DNA ligase. The present invention provides a DNA vector used to identify gene expression patterns in an mRNA population, for example, for use in the methods of the present invention. Preferably, the DNA vector includes an insertion site; a restriction endonuclease recognition sequence, Sequence A, located 5' to the insertion site wherein the restriction endonuclease that recognizes Sequence A has a cleavage site, Sequence B, located 3' to the Sequence A; and a restriction endonuclease recognition sequence, Sequence C, located 5' to or overlapping with the Sequence A. Sequence A can be the same as the second restriction endonuclease recognition sequence used in the methods described herein. Sequence C can be the same as the third restriction endonuclease recognition sequence used in the methods described herein. The insertion site of the vector preferably is compatible with the ends of the cDNA inserts. The sequences may also be recognized by restriction endonucleases having compatible ends with the priming and the first restriction endonucleases used to obtain the cDNA insert, as long as the use of the endonucleases and the insertion of the cDNA inserts maintains the integrity of a cleavage site at the first restriction endonuclease site. If only one of the ends is compatible, the cDNA insert can be inserted using blunt end ligation at one of the ends. Thus, in preferred aspects, the insertion site has two ends, wherein the first end is compatible with a first insertion restriction endonuclease cleavage site and the second end is compatible with a second insertion restriction endonuclease cleavage site. The first insertion restriction endonuclease cleavage site is preferably compatible with the first restriction endonuclease cleavage site. The second insertion restriction endonuclease cleavage site is preferably compatible with the second restriction nuclease cleavage site. In a preferred embodiment, the vector includes Sequence A which is recognized by a Type IIs restriction endonuclease such that the cleavage site, Sequence B, is 3' to Sequence A; a restriction endonuclease recognition sequence, Sequence C, located 5' to or overlapping with Sequence A; a restriction endonuclease cleavage site, Sequence D, located 3' to Sequence A and 5' to Sequence B, wherein Sequence D can be cleaved by a restriction endonuclease that recognizes less than six bases; and a restriction endonuclease cleavage site, Sequence E, which can be cleaved by a restriction endonuclease that recognizes more than six bases. Most preferably, the DNA vector is the vector depicted in Figure 1. In other preferred embodiments, the invention provides DNA constructs that include DNA vectors described herein which include DNA inserts at the insertion sites. In one embodiment, the DNA vector of the present invention further comprises a cDNA insert inserted at the insertion site, wherein Sequence B is within the cDNA insert.

In other embodiments, the invention provides methods for isolating a gene. The methods include cleaving double-stranded cDNAs with a first restriction endonuclease to obtain cDNA inserts. The cDNA inserts are inserted into the insertion sites of cloning vectors to obtain a DNA construct. The cloning vectors typically include a second restriction endonuclease recognition sequence 5' to the insertion site such that digestion of the DNA construct with the second restriction endonuclease cleaves the DNA construct at a site within the cDNA insert, and a third restriction endonuclease recognition sequence 5' to or overlapping with the second restriction endonuclease recognition sequence. The DNA constructs are amplified, isolated, and then digested with the second and the third restriction endonuclease to obtain tags. The tag that comprises a portion of the sequence of the gene to be isolated is identified and the gene is isolated. In preferred aspects, the gene to be isolated is determined by comparing the nucleotide sequence of a tag with known nucleotide sequences which can be obtained from any source, such as sequence databases, e.g., GenBank.

In other aspects of the invention there is provided a method for identifying gene expression patterns in an mRNA population. The method includes preparing double-stranded cDNAs from an mRNA population using a primer, e.g., an oligo dT sequence linked at the 5' end of the oligo dT sequence to a cleavage site for a "priming" restriction endonuclease, and cleaving the double-stranded cDNAs with a first restriction endonuclease, which cleaves at a site within the cDNA sequence and not within the primer, to obtain cDNA inserts. The cDNA inserts are inserted into the insertion sites of cloning vectors to obtain DNA constructs, wherein the cloning vectors include a second restriction endonuclease recognition sequence 5' to the insertion site such that digestion of the DNA constructs with the second restriction endonuclease cleaves the DNA constructs at sites within the cDNA inserts. DNA constructs are amplified, e.g., in a suitable host cell, e.g., *E. coli*, isolated, and then digested with the second restriction endonuclease to obtain a linearized DNA molecule having a 3' overhang sequence. The linearized DNA molecule is annealed to an adapter sequence. The adapter sequence includes a double-stranded oligodeoxynucleotide sequence comprising a first restriction endonuclease recognition sequence and a 3' underhang sequence compatible with the 3' overhang sequence of the linearized DNA molecule. Annealing of the adapter to the linearized DNA molecule results in a ligation product flanked by first restriction endonuclease restriction sites. The ligation product is digested with the first restriction endonuclease to obtain tags. The nucleotide sequence of the tags is obtained to identify gene expression patterns in the mRNA population. A preferred aspect of the second embodiment is outlined in Figure 3.

In preferred aspects, the nucleotide sequence of the tags is obtained by ligating the tags to obtain ligated tag arrays of at least about 10 tags, inserting the ligated tag arrays into a sequencing vector, and sequencing the ligated tag arrays. In a preferred embodiment, the adapter is about 10 to about 15 base pairs in length and it includes a degenerate sequence, e.g., two base pairs in length, as the 3' underhang and the linearized DNA molecule includes a degenerate sequence, e.g., two base pairs in length, as the 3' overhang.

The invention also provides a method for identifying gene expression patterns in a population of mRNA. The method includes preparing a population of double-stranded cDNA from a first population of mRNA obtained from a first biological sample, using a primer, e.g., an oligo dT sequence, covalently linked to an affinity capture label, e.g., biotin, and cleaving the double-stranded cDNA with a punctuating restriction endonuclease, which cleaves at a site within the cDNA and not within the primer, to obtain a population of cDNA inserts linked to the affinity capture label. The cDNA inserts are captured by capturing the affinity capture label with an affinity capture device, e.g., magnetic beads covalently coupled to streptavidin, to obtain a population of captured cDNA inserts. The captured cDNA insert is then annealed and ligated to a first adapter which includes a double-stranded oligodeoxynucleotide sequence comprising a 5' overhang sequence compatible with a first vector insertion site, a second restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a punctuating restriction endonuclease site, to obtain a first ligation product. Cleavage of the first ligation product with a second restriction endonuclease, e.g., a Type IIs restriction endonuclease, releases the ligation product separated from the affinity capture label, wherein the released ligation product comprises a punctuating endonuclease restriction site adjacent to a cDNA sequence and a 3' overhang sequence. The released ligation product is annealed and ligated with a second adapter which includes a double-stranded oligodeoxynucleotide sequence comprising a 5' underhang sequence compatible with a second vector insertion site and a 3' underhang sequence compatible with the 3' overhang sequence of the released ligation product. This annealing step yields a second ligation product which includes a 5' sequence compatible with a first vector insertion site, cDNA sequence flanked by punctuating endonuclease restriction sites, and a 3' sequence compatible with a second vector insertion site. The second ligation product is then inserted into a cloning vector at a first vector insertion site and a second vector insertion site to obtain a DNA construct. The DNA construct is amplified, e.g., in a suitable host cell, e.g., *E.coli*, isolated and digested with the punctuating restriction endonuclease to obtain tags. The nucleotide sequence of the tags is obtained to identify gene expression in the first biological sample.

In preferred aspects, the nucleotide sequence of the tags is obtained by ligating the tags to obtain ligated tag arrays of at least about 10 tags, more preferably of at least about 40 tags, wherein each tag in the tag array is adjacent to a punctuating restriction endonuclease recognition site, inserting the ligated tag arrays into a sequencing vector, sequencing the ligated tag arrays, and comparing sequences of the tag array to known gene sequences. Preferably, the method further includes the step of isolating a gene sequence that hybridizes to a tag. In a preferred embodiment, the second restriction endonuclease cleavage site is located about 16 nucleotides 3' of its recognition sequence. In one embodiment, the first adapter includes the second restriction endonuclease recognition site located 5' to sequence which is compatible with the punctuating restriction endonuclease site. In another embodiment, the released ligation product includes a 3' overhang of two nucleotides in length, and the second adapter includes a 3' underhang sequence comprising two nucleotides of degenerate sequence. In another embodiment, 5' overhang sequence compatible with the first vector insertion site includes a restriction endonuclease recognition sequence of at least eight bases, e.g., a NotI recognition sequence, and the 5' underhang sequence compatible with the second vector insertion site is an EcoRI recognition sequence. In yet another embodiment, 5' overhang sequence compatible with the first vector insertion site includes an EcoRI recognition sequence, and the 5' underhang sequence compatible with the second vector insertion site is a NotI recognition sequence. In a preferred embodiment, the first adapter is about 15 to about 25 base pairs in length and the second adapter includes a degenerate sequence, e.g., two base pairs in length, as the 5' underhang insert space. Preferably, the second restriction endonuclease recognition site of the first adapter is located 5' to the sequence which is compatible to the punctuating endonuclease restriction site. In another embodiment, the released ligation product is annealed to a mixture of 16 different adapters each having a different degenerate sequence. Preferably, the 3' overhang of the released ligation product is two base pairs in length. In preferred embodiments, the ligated tag arrays include at least about 30 tags, preferably at least about 50 tags, more preferably at least about 100 tags, and most preferably at least about 200 tags. In one embodiment, the cloning vector lacks punctuating endonuclease restriction sites.

In a preferred embodiment, the method further includes preparing an oligonucleotide probe comprising a nucleotide sequence of a tag; and probing a cDNA library with the olignucleotide probe to determine a frequency of expression of a gene which comprises the tag. In another embodiment, the method further includes repeating the method of the third embodiment using a second population of mRNA from a second biological sample; and comparing gene expression of the first population of mRNA with

gene expression of the second population of mRNA to determine differences in gene expression between the first biological sample and the second biological sample. Preferably, the method further includes identifying a gene that is expressed at a first level in the first population of mRNA and is expressed at a second level in the second population of mRNA; and isolating the gene from a cDNA library. In a preferred embodiment, the first biological sample is cells or tissue obtained from a normal non-diseased organism, and the second biological sample is cells or tissue obtained from an organism having a disease or disorder. In another preferred embodiment, the first biological sample is cells or tissue obtained from an organism at a first stage of development, and the second biological sample is cells or tissue obtained from an organism at a second stage of development.

In a preferred aspect, a method is provided for identifying gene expression patterns in an mRNA population. The method includes preparing a population of double-stranded cDNA from a first population of mRNA obtained from a first biological sample, using a primer comprising a 5' oligo dT sequence covalently linked at a 3' end to a biotin label, and cleaving the double-stranded cDNA with a Sau3A restriction endonuclease to obtain a population of cDNA inserts linked to biotin label. The cDNA inserts are captured by capturing biotin label with magnetic beads covalently coupled to streptavidin, to obtain a population of captured cDNA inserts. The captured cDNA insert is then annealed and ligated to a first adapter which includes a double-stranded oligodeoxynucleotide sequence comprising a 5' overhang sequence compatible with a NotI insertion site, a BsgI restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a Sau3A restriction site, to obtain a first ligation product. Cleavage of the first ligation product with BsgI relaeases the ligation product separated from the biotin label, wherein the released ligation product comprises a Sau3A restriction site adjacent to a cDNA sequence and a 3' overhang sequence. The released ligation product is annealed and ligated with a second adapter which includes a double-stranded oligodeoxynucleotide sequence comprising a 5' underhang sequence compatible with an EcoRI insertion site and a 3' underhang degenerates sequence compatible with the 3' overhang sequence of the released ligation product. This annealing step yields a second ligation product which includes a 5' sequence compatible with a NotI insertion site, cDNA sequence flanked by Sau3A restriction sites, and a 3' sequence compatible with an EcoRI insertion site. The second ligation product is then inserted into a cloning vector at a NotI insertion site and an EcoRI insertion site to obtain a DNA construct. The DNA construct is amplified, e.g., in a suitable host cell, e.g., *E.coli*, isolated and digested with Sau3A to obtain tags which are then ligated to obtain ligated tag arrays of

about 30-60 tags. The nucleotide sequence of the tags is obtained to identify gene expression in the first biological sample.

In a related aspect, the present invention provides a DNA vector used to identify gene expression patterns in an mRNA population, for example, for use in the methods of the present invention. The DNA vector includes an insertion site and lacks punctuating endonuclease restriction sites, and preferably also includes a cDNA insert which includes at least one punctuating endonuclease restriction site, e.g., a Sau3A restriction site. Preferably, the insertion site has two ends, wherein the first end is compatible with a first insertion restriction endonuclease cleavage site and the second end is compatible with a second insertion restriction endonuclease cleavage site. Preferably, the first insertion restriction endonuclease cleavage site includes at least eight bases, e.g., the cleavage site is a NotI cleavage site, and the second restriction endonuclease cleavage site comprises at least six bases, e.g., the cleavage site is an EcoRI cleavage site. In other preferred embodiments, the invention provides DNA constructs, including the DNA vector, e.g., the TALESTB vector described herein, and further including a DNA insert between a NotI and an EcoRI insertion site.

The present invention provides a method for determining the frequency of gene expression in an mRNA population. The method includes preparing the DNA constructs including the cDNA inserts of the present invention to obtain a cDNA library. The method further includes preparing an oligonucleotide probe comprising a tag sequence identified according to the methods of the invention and probing the cDNA library with the oligonucleotide probe including the tag sequence to determine the frequency of expression of a gene which includes the tag sequence. Other embodiments include methods for isolating a gene that is expressed at different levels in a first mRNA population compared to a second mRNA population. These methods include identifying a gene expression pattern from a first mRNA population and identifying a gene expression pattern from a second additional mRNA population according to the present invention. The gene expression patterns so obtained are compared to detect differences in gene expression between the mRNA populations. A gene that is expressed at a different level in the first mRNA population compared to the second mRNA population can then be identified and isolated. Another embodiment is a method for detecting a difference in gene expression between two or more mRNA populations. The method includes identifying a gene expression pattern from a first mRNA population and from at least one additional mRNA population according to the methods of the present invention. The gene expression patterns so obtained are compared, thereby detecting differences in gene expression between the mRNA populations. In preferred aspects, the first mRNA population is obtained from a normal cell or tissue and the additional

mRNA population is obtained from a cell or tissue from a target organism having a disease or disorder. In other preferred aspects, the mRNA populations are obtained from cells or tissues at different developmental stages. In yet other preferred aspects, the mRNA populations are obtained from cells derived from different tissues or organs of the same target organism or the mRNA populations are obtained from different target organisms. Another embodiment provides methods for detecting the presence of a disease in a target organism. These methods include identifying a gene that is expressed differently in a normal cell or tissue than in a cell or tissue from a target organism having a disease or disorder according to the methods of the present invention and isolating the tag sequence of the gene. An mRNA population obtained from a first target organism and an mRNA population obtained from a second normal or diseased target organism can be probed with the tag sequence to determine the level of expression of the gene. The level of expression of the gene in the first target organism is compared with the level of expression of the gene in the second target organism to detect the presence of a disease in the first target organism. Yet another aspect provides a method for screening for the effects of a drug on a cell or tissue. The methods of the invention can be used to compare mRNA gene expression patterns in cells and tissues that have been treated with a drug versus cells and tissues that have not been treated with a drug. The cells or tissues can be from normal target organisms and the side effects of a drug can be tested. Alternatively, the cells or tissues can be from diseased target organisms with particular disorders to determine whether the drug can change the gene expression profile in the diseased cells.

In another preferred aspect, the invention provides a method for isolating a differentially expressed gene. The method includes obtaining the nucleotide sequence of ligated tag arrays obtained from a first cell type or tissue and from a second cell type or tissue according to the methods of the invention and comparing the frequency of expression of the individual tag sequences of the first and second cell types or tissues. Differentially expressed tag sequences in the first cell type or tissue compared to the second cell type or tissue are identified and a gene corresponding to the differentially expressed tag sequences can then be identified. In preferred aspects, the genes are identified by searching a database of RNA or DNA sequences for the differentially expressed tag sequence. Alternatively, the genes are identified by probing a cDNA library with a probe comprising the differentially expressed tag sequences.

In yet another embodiment, the invention provides kits for use in the methods described herein, e.g., in identifying gene expression patterns in mRNA populations or in isolating a gene that is differentially expressed. In a preferred embodiment, a kit for use in identifying gene expression patterns in an mRNA population includes a DNA

vector, e.g., the TALEST vector described herein, a primer comprising about 7 to 40 T residues, a first restriction endonuclease that recognizes the Sequence A and cleaves DNA at the Sequence B, and a second restriction endonuclease that recognizes Sequence C. In another embodiment, a kit for use in identifying gene expression patterns in an mRNA population includes a DNA vector, e.g., the TALESTB vector described herein, e.g., a DNA vector comprising a NotI insertion site, an EcoRI insertion site, and one or fewer Sau3A restriction endonuclease recognition sites, a primer comprising about 7 to 40 T residues, a first adapter which includes a double-stranded oligodeoxynucleotide sequence including a second restriction endonuclease, e.g., a Type IIs restriction endonuclease, recognition sequence, a 5' overhang sequence compatible with a first vector insertion site, e.g., a NotI insertion site, and a 5' underhang sequence compatible with a pancutating endonuclease restriction site, e.g., a Sau3A restriction site, and a second adapter which includes a double-stranded oligodeoxynucleotide sequence including a 3' underhang sequence, e.g., a degenerate sequence, and a 5' underhang sequence compatible with a second vector insertion site, e.g., an EcoRI insertion site.

Brief Description of the Drawings

Figure 1 depicts the TALEST vector which can be used in the first (TALEST) embodiment of the invention.

Figures 2A and 2B depict a schematic representation of the first (TALEST) embodiment of the present invention.

Figure 3 depicts a schematic representation of the second (TALESTA) embodiment of the present invention.

Figure 4 depicts a schematic representation of the TALEST method.

Figure 5 depicts a schematic representation of the the third (TALESTB) embodiment of the present invention.

Figure 6 depicts a schematic representation of another embodiment of the present invention.

Detailed Description of the Invention

The present invention provides novel methods for identifying gene expression patterns in mRNA populations. The methods are useful for determining differential gene expression among various cells or tissues, including cells or tissues of a target organism. The invention also provides methods of determining the frequency of gene expression in mRNA populations, thus providing a method of comparing gene expression frequency among various cells or tissues. The present invention also provides methods for isolating genes corresponding to tag sequences identified according to the methods of

the present invention. Furthermore, sequences that are identified according to the present invention may be used to diagnose the presence of disease.

In order to fully understand gene expression patterns of a particular cell lineage, it is necessary to know not only which genes are expressed by the cell, but also the frequencies or rates at which they are expressed. The methods of the present invention provide novel methods for identifying gene expression patterns in cells and tissues and methods for determining the frequency of gene expression in cells and tissues in a simple and reproducible manner that does not require the use of PCR or other methods that may limit the reproducibility of the assays. Furthermore, the methods of the present invention are not limited by the ability of the researcher to synthesize numerous oligonucleotide primers to correspond to the huge variety of mRNA sequences. By obtaining the RNA sequence tags according to methods of the present invention, the frequency of gene expression can be determined merely by analyzing the frequency of cDNA expression in the cDNA library prepared during the process of producing the tags.

At least three different embodiments of the present invention are described in detail in the subsections below.

The TALEST Embodiment

The first or TALEST (***t***andem ***a***rrayed ***l***igation of ***e***xpressed ***s***equence ***t***ags) embodiment includes a method for identifying gene expression patterns in an mRNA population. The method includes preparing double-stranded cDNAs from an mRNA population using a primer, and cleaving the double-stranded cDNAs with a first restriction endonuclease, which cleaves at a site within the cDNA sequence and not within the primer, to obtain a population of cDNA inserts. A cDNA insert is inserted into the insertion iste of a cloning vector to obtain a DNA construct, wherein the cloning vector includes a second restriction endonuclease recognition sequence 5' to the insertion site such that digestion of the DNA construct with the second restriction endonuclease cleaves the DNA construct at a site within the cDNA insert, and a third restriction endonuclease recognition sequence 5' to or overlapping with the second restriction endonuclease recognition sequence. DNA constructs are amplified, isolated, and digested with the second and the third restriction endonuclease to obtain tags. The nucleotide sequence of the tags is obtained to identify gene expression patterns in the mRNA population.

Herein, "gene" refers to a unit of inheritable genetic material found in a chromosome, such as in a human chromosome. Each gene is composed of a linear chain of deoxyribonucleotides which can be referred to by the sequence of nucleotides forming the chain. Thus, "sequence" is used to indicate both the ordered listing of the

nucleotides which form the chain, and the chain which has that sequence of nucleotides. (The term "sequence" is used in the same way in referring to RNA chains, linear chains made of ribonucleotides.) The gene includes regulatory and control sequences, sequences which can be transcribed into an RNA molecule, and may contain sequences with unknown function. Some of the RNA products (products of transcription from DNA) are messenger RNAs (mRNAs) which initially include ribonucleotide sequences (or sequence) which are translated into a polypeptide and ribonucleotide sequences which are not translated. The sequences which are not translated include control sequences, introns and sequences with unknown function. It should be recognized that small differences in nucleotide sequence for the same gene can exist between different persons, or between normal cells and cancerous cells, without altering the identity of the gene.

Herein, the term "gene expression pattern" means the set of genes of a specific tissue or cell type that are transcribed or "expressed" to form RNA molecules. Which genes are expressed in a specific cell line or tissue will depend on factors such as tissue or cell type, stage of development of the cell, tissue, or target organism and whether the cells are normal or transformed cells, such as cancerous cells. For example, a gene may be expressed at the embryonic or fetal stage in the development of a specific target organism and then become non-expressed as the target organism matures. Alternatively, a gene may be expressed in liver tissue but not in brain tissue of an adult human. The list of factors affecting expression and the examples are not exhaustive; and are intended only as illustration.

Preferably, the primer used to prime cDNA synthesis consists of an oligo dT sequence linked at the 5' end of the oligo dT sequence to a cleavage site for a "priming" restriction endonuclease. The oligo dT sequence is preferably about 7 to 40 T residues in length, more preferably the oligo dT sequence is about 15 to 30 T residues in length. Most preferably, the oligo dT sequence is about 19 T residues in length. In order to maximize the number of mRNAs that can be identified using the methods of the present invention, the priming restriction endonuclease should recognize very few sequences. Thus, preferred priming restriction endonucleases recognize sequences consisting of more than six bases and are known to those skilled in the art. The priming restriction endonuclease is preferably one that recognizes an eight-base palindromic sequence. More preferably, the priming restriction endonuclease recognizes a sequence comprising at least one CG dinucleotide. Most preferably, the priming restriction endonuclease is NotI.

Herein, the term "first restriction endonuclease" refers to a restriction endonuclease which recognizes a sequence consisting of less than six base pairs in DNA,

preferably it recognizes a four base pair sequence in DNA. Examples of a first restriction endonuclease include, but are not limited to, MboI, Sau3A, MspI, AluI, BstUI, DpnII, HaeIII, HhaI, HinPI, MseI, NlaIII, RmaI, and TaqI.

Herein, the term "cDNA insert" refers to a cDNA sequence that can be inserted into a vector. Typically, the cDNA insert is about 250, 300 or 350 base pairs in length. Preferably, the cDNA insert includes a polyA tail.

Herein, a "vector" means an agent into which DNA of this invention can be inserted by incorporation into the DNA of the agent. Thus, examples of classes of vectors can be plasmids, cosmids, and viruses (*e.g.*, bacteriophage). Typically, the agents are used to transmit the DNA of the invention into a host cell (*e.g.*, bacterium, yeast, higher eukaryotic cell). A vector can be chosen based on the size of the insert desired, as well as based on the proposed use of the vector. For preservation of a specific DNA sequence (e.g., in a cDNA library) or for producing a large number of copies of the specific DNA sequence, a cloning vector can be used. For transcription of RNA or translation to produce an encoded polypeptide, an expression vector can be used. Following transfection of a cell, all or part of the vector DNA, including the insert DNA, can be incorporated into the host cell chromosome, or the vector can be maintained extrachromosomally.

Those skilled in the art will recognize that the vector comprising the cDNA insert or fragment, (i.e., the DNA construct), can be amplified using any method known in the art. Preferably, the construct is amplified in a host cell such as, but not limited to, *E. coli* by first transforming *E. coli* with the construct, growing the transformed cells, and isolating the amplified vector from the grown cells.

As used herein, the term "second restriction endonuclease" refers to a restriction endonuclease which cleaves downstream or 3' from its own recognition sequence. Preferred second restriction endonucleases are Type IIs restriction endonucleases. Examples of Type IIs restriction endonucleases which can be used in the methods of the present invention include BsgI, FokI, AccBSI, AceIII, AciI, AclWI, AlwI, Alw26I, AlwXI, Asp26HI, Asp27HI, Asp35HI, Asp36HI, Asp40HI, Asp50HI, AsuHPI, BaeI, BbsI, BbvI, BbvII, Bbv16II, Bce83I, BcefI, BcgI, Bco5I, Bco116I BcoKI, BinI, Bli736I, BpiI, BpmI, Bpu10I, BpuAI, Bsal, BsaMI, Bsc9II, BscAI, BscCI, BseII, Bse3DI, BseNI, BseRI, BseZI, BsiI, BsmI, BsmAI, BsmBI, BsmFI, Bsp24I, Bsp423I, BspBS3II, BspIS4I, BspKT5I, BspLU11III, BspMI, BspPI, BspST5I, BspTS514I, BsrI, BsrBI, BsrDI, BsrSI, BssSI, Bst11I, Bst71I, Bst2BI, BstBS32I, BstD102I, BstF5I, BstTS5I, Bsu6I, CjeI, CjePI, Eam1104I, EarI, Eco31I, Eco57I, EcoA4I, EcoO44I, Esp3I, FauI, GdiII, GsuI, HgaI, HphI, Ksp632I, MboII, MlyI, MmeI, Mn1I, Mva1269I, PhaI, PieI, RleAI, SapI, SfaNI, SimI, StsI, TaqII, TspII, TspRI, Tth111II, and VpaK32I.

A "third restriction endonuclease", as used herein, refers to a restriction endonuclease which cleaves 3' from its own recognition sequence. Preferred third restriction endonucleases are Type IIs restriction endonucleases.

As used herein, the term "isolating" refers to a method by which the DNA construct is separated from the reagents used in amplification. Preferably, the DNA construct is substantially free of amplification buffer, primers, cellular material, culture medium or gel material.

The term "tag" refers to a nucleotide sequence which includes a sufficient number of base pairs such that it uniquely defines a cDNA sequence. Typically, for a tag to uniquely identify a eukaryotic cDNA sequence, the tag includes at least about eight base pairs in length. In a preferred embodiment, the tag is at least about 10, 12 or 14 base pairs in length. Once the tags of the present invention are obtained, they are preferably ligated to produce tag arrays, e.g., at least two tags ligated in series. Preferably, the tag arrays include at least 10, more preferably at least 20, still more preferably at least 30, yet more preferably at least 40, and even more preferably at least 50 or more, e.g., 100, 150, 200 or more, tags. To sequence the tags in the tag arrays, the arrays can be inserted into a sequencing vector and the sequenced.

The present invention also provides DNA vectors and kits for use in the TALEST embodiment. A preferred DNA vector includes an insertion site; a restriction endonuclease recognition sequence (Sequence A), located 5' to the insertion site wherein the restriction endonuclease has a cleavage site (Sequence B), located 3' to the Sequence A; and a restriction endonuclease recognition sequence (Sequence C), located 5' to or overlapping with the Sequence A. Sequence A can be the same as the second restriction endonuclease recognition sequence used in the methods of the present invention described herein. Sequence C can be the same as the third restriction endonuclease recognition sequence used in the methods described herein. A preferred kit for use in identifying gene expression patterns in an mRNA population includes a DNA vector, e.g., the TALEST vector described herein, a primer comprising about 7 to 40 T residues, a first restriction endonuclease that recognizes the Sequence A and cleaves DNA at the Sequence B, and a second restriction endonuclease that recognizes Sequence C.

An overview of the first or TALEST embodiment of the present invention is presented in Figures 2 and 4. Although the overview presented in Figures 2 and 4 and described herein provides a detailed description of the invention using particular restriction endonucleases, and a defined vector, it is well known to those of skill in the art that other restriction endonucleases can be selected and other methods of molecular biology, such as those described in Sambrook J. et al., "Molecular Cloning: A laboratory Manual", Second Ed. (Coldspring Harbor Laboratory Press, Cold Spring Harbor, New

York, 1989, Volume 1, Chapter 7), can be used to practice the present invention and this invention is not limited to the detailed examples presented herein.

Polyadenylated mRNA is first isolated from the cell population of interest using standard procedures. The mRNA is then converted to cDNA using reverse transcriptase by priming the mRNA with an oligo dT sequence that has a rare cutting enzyme site (e.g., NotI) at its 5' end. The sequence of a suitable primer that can be used to prime cDNA synthesis is 5'TTTTTTTTTTTTTTTTTTTTCGCCGGGCGCATG 3' (SEQ ID NO:3), which comprises an oligo dT sequence linked to a NotI endonuclease recognition sequence. The first strand cDNA is converted to double-stranded cDNA using RNAase H and DNA polymerase 1. The double-stranded cDNA is then digested with two different restriction enzymes (e.g., NotI and MboI). The use of two restriction enzymes allows the cDNA to be directionally cloned into the TALEST vector depicted in Figure 1.

The TALEST vector contains a Not I recognition site and a Bam HI recognition site, which, when cleaved with Bam HI endonuclease produces ends compatible with MboI endonuclease digested DNA. MboI has a four base recognition sequence (GATC) which occurs in eukaryotic DNA on an average of once every 256 base pairs. Thus, the average size of the cloneable NotI/MboI cDNA fragment is approximately 300 base pairs including the portion of the poly A tail that has been cloned. When the cDNA is cloned into the TALEST vector, a cDNA library is formed that is representative of virtually all the expressed genes in the cell.

The library is prepared in a directional orientation such that the 5' terminus of every cDNA in the library always begins with the MboI recognition sequence, the GATC sequence, which in turn is derived from the 3' most MboI site found in the gene. The library is then amplified by transforming the plasmid into a host cell and allowing the bacteria to grow.

The TALEST vector has a BsgI restriction endonuclease site located immediately 5' to the GAT sequence that begins every cDNA. BsgI is a Type IIs restriction endonuclease which recognizes a defined sequence (GTGCAG) but cleaves the DNA approximately 16 bases "downstream" (3') from the recognition sequence. Thus, cleavage of the TALEST vector with BsgI linearizes the circular plasmid by cleaving the inserted cDNA 12 bases downstream from the GATC start sequence on the sense strand, and 10 bases on the antisense strand. Because BsgI leaves a 3' "overhang," the unpaired two bases on the sense strand are removed using T4 DNA polymerase to generate blunt ends.

Nine bases upstream from the BsgI site is a second Type IIs restriction site, FokI. This enzyme recognizes the 5-base sequence GGATG but cleaves 9 bases downstream

(3') on the sense strand, and 13 bases downstream on the antisense strand. When the resultant fragment is subjected to treatment with T4 DNA polymerase, a blunt-ended 15 base "tag" is generated with the sequence: GGATCNNNNNNNNNN (SEQ ID NO:4).

Alternatively, PmlI can be used as the second restriction site. This site is convenient because its recognition sequence (CACGTG) overlaps that of BsgI and it cleaves both the sense and antisense strands of the DNA at the same place leaving blunt ends. Digestion of the BsgI linearized plasmid with Pml1 cleaves off the 20 base blunt ended fragments with the sequence GTGCAGGATCNNNNNNNNNN (SEQ ID NO:5) where the first six bases are derived from the TALEST vector and the next 14 (GATCNNNNNNNNNN (SEQ ID NO:6)) are derived from the cDNA.

When the entire amplified cDNA library is digested with BsgI and FokI, a 20 base pair fragment is excised which consists of a mixture of "tags," each of which differs in the sequence of the final ten bases and each of which uniquely mark a single expressed gene. With ten bases of unknown sequence, there are 4^{10} or 1,048,576 possible different tag sequences. This number exceeds by approximately five-fold the number of expressed genes in the human genome in all tissues.

The tags are mixed together and subjected to enzymatic treatment with DNA ligase in order to generate tandem arrays of about 30-60, preferably about 40-50 tags in a single molecule. The arrays are then cloned into a sequencing vector and subjected to automated DNA sequence analysis. When the arrays are analyzed, individual tags are recognized because they are separated from each other by the defined punctuation sequence, GGATC (containing the MboI recognition sequences) or its reverse complement depending on the random sense or antisense or orientation of the tag during ligation.

Each tag begins with the defined GGATC sequence derived from the 3' most MboI site in the original cDNA, and has ten additional bases of unknown sequence that uniquely marks one of the expressed genes in the cell population under study. The presence of the GGATC start sequence effectively provides five bases of additional identifying information, and localizes the information to a particular site within the tagged gene. Thus, in effect, 15 bases of sequence are known for each mRNA that has been copied into cDNA and is analyzed in the present method.

The TALESTA Embodiment

The second or TALESTA embodiment includes another method for identifying gene expression patterns in an mRNA population. The method includes preparing double-stranded cDNAs from an mRNA population using a primer, e.g., an oligo dT sequence linked at the 5' end of the oligo dT sequence to a cleavage site for a "priming"

restriction endonuclease, and cleaving the double-stranded cDNAs with a first restriction endonuclease, which cleaves at a site within the cDNA sequence and not within the primer, to obtain a population of cDNA inserts. A cDNA insert is inserted into the insertion site of a cloning vector to obtain a DNA construct, wherein the cloning vector includes a second restriction endonuclease recognition sequence 5' to the insertion site such that digestion of the DNA construct with the second restriction endonuclease cleaves the DNA construct at a site within the cDNA insert. DNA constructs are amplified, e.g., in a suitable host cell, e.g., *E. coli*, isolated, and then digested with the second restriction endonuclease to obtain a linearized DNA molecule having a 3' overhang sequence. The linearized DNA molecule is annealed and ligated to an adapter sequence. The adapter sequence includes a double-stranded oligodeoxynucleotide sequence comprising a first restriction endonuclease recognition sequence and a 3' underhang sequence compatible with the 3' overhang sequence of the linearized DNA molecule. Annealing and ligating of the adapter results in a linearized DNA molecule ligation product comprising cDNA flanked by first restriction endonuclease restriction sites. The ligation product is digested with the first restriction endonuclease to obtain tags. The nucleotide sequence of the tags is obtained to identify gene expression patterns in the mRNA population.

Herein, the term "adapter" refers to a double-stranded oligodeoxynucleotide sequence, wherein the sequence of the top strand is in a 5' to 3' orientation and the sequence of the bottom strand is in a 3' to 5' orientation with respect to each other.

Herein, the term "3' underhang" refers to a single-stranded sequence located at the 3' end of the bottom strand of an adapter.

Herein, the term "3' overhang" refers to a single-stranded sequence located at the 3' end of the top strand of an adapter.

The present invention also provides DNA vectors and kits for use in the TALESTA embodiment. A preferred kit for use in identifying gene expression patterns in an mRNA population includes a DNA vector, e.g., a DNA vector which includes a punctuating restriction endonuclease recognition sequence adjacent (3') to a degenerate sequence which can be digested to leave a degenerate overhang.

An overview of the second or TALESTA embodiment of the present invention is presented in Figure 3. Although the overview presented in Figure 3 and described herein provides a detailed description of the invention using particular restriction endonucleases, and a defined vector, it is well known to those of skill in the art that other restriction endonucleases can be selected and other methods of molecular biology, such as those described in Sambrook J. et al., "Molecular Cloning: A laboratory Manual", Second Ed. (Coldspring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989,

Volume 1, Chapter 7), can be used to practice the present invention and this invention is not limited to the detailed examples presented herein.

Polyadenylated mRNA is first isolated from the cell population of interest using standard procedures. The mRNA is then converted to cDNA using reverse transcriptase by priming the mRNA with an oligo dT sequence with an affinity capture label (such as biotin) at its 5' end. The first strand cDNA is converted into double stranded cDNA using RNAse H and DNA Polymerase I using standard procedures. The double stranded cDNA is then cleaved with a punctuating restriction endonuclease having a four base-pair recognition sequence. The "punctuating restriction endonuclease" refers to an endonuclease that cleaves cDNA leaving a recognition sequence that will be present at the 5' end of every tag sequence, so that when the tag sequences are concatenated, the recognition sequence will be present at each end of the tag sequence, thus serving as punctuating sequences between the concatenated cDNA sequences. The 3' fragment containing the affinity capture label is purified using an affinity capture device (such as streptavidin conjugated magnetic beads). This captured cDNA fragment is then annealed to an adapter including a double stranded oligodeoxynucleotide sequence including a Type IIs restriction endonuclease recognition sequence, a 5' overhang sequence compatible with a punctuating endonuclease restriction site to form a first ligation product. This first ligation product is then cleaved with a Type IIs restriction endonuclease to release the ligation product from the affinity capture device wherein the released ligation product includes the punctuating endonuclease recognition sequence adjacent to a cDNA insert derived sequence wherein the cDNA-derived sequence includes a degenerate 1 or 2-base 3' overhang sequence. The released ligation product is then ligated using standard techniques into a degenerate cloning vector such that the 5' restriction endonuclease recognition sequence of the ligation product is compatible with overhangs generated by restriction digest of the vector, and the 3' site is compatible with overhangs introduced by digestion of the degenerate vector with the same Type IIs restriction endonuclease used to remove the fragment from the affinity capture device. The degenerate vector also contains the punctuating restriction endonuclease recognition sequence immediately adjacent (3') to the degenerate overhang. The DNA construct is then transformed into competent bacteria and amplified by standard techniques to generate a tag library. After amplification, the tags are released by digesting the vector DNA with the punctuating restriction endonuclease.

The TALESTB Embodiment

The third or TALESTB embodiment includes yet another method for identifying gene expression patterns in an mRNA population. The method includes preparing

double-stranded cDNAs from an mRNA population using a primer having an affinity capture label and cleaving the double-stranded cDNAs with a punctuating endonuclease having a four base pair recognition sequence, to obtain cDNA inserts. A 3' cDNA insert having the affinity capture label is captured on an affinity capture device to obtain a captured cDNA insert. The captured cDNA insert is annealed to a first adapter including a double-stranded oligodeoxynucleotide sequence including a second restriction endonuclease (e.g., a Type IIs restriction endonuclease), recognition sequence, a 5' overhang sequence compatible with a first vector insertion site and a 5' underhang sequence compatible with a punctuating endonuclease restriction site, to obtain a first ligation product. The first ligation product is cleaved with a second restriction endonuclease (e.g., a Type IIs restriction endonuclease), to release the ligation product from the affinity capture device, wherein the released ligation product includes the punctuating endonuclease restriction site adjacent to a cDNA insert derived sequence wherein the cDNA derived sequence includes a 3' overhang sequence. The released ligation product is annealed with a second adapter, i.e., a double-stranded oligodeoxynucleotide sequence that includes a 3' underhang sequence compatible with the 3' overhang sequence of the ligation product and a 5' underhang sequence compatible with a second vector insertion site, to obtain a second ligation product. The second ligation product includes a cDNA derived sequence flanked by the punctuating endonuclease restriction sites, the first vector insertion site on a 5' end of the product, and the second vector insertion site on a 3' end of the product. The second ligation product is inserted into the insertion sites of a cloning vector to obtain a DNA construct. The DNA construct is amplified, isolated, and digested with the punctuating endonuclease to obtain tags. The nucleotide sequence of the tags is obtained to identify gene expression patterns in the mRNA population.

Herein, the term "affinity capture label" refers to a moiety which can be linked to or included within a primer and which is capable of interacting with (e.g., binding to) a capture moiety, e.g., an affinity capture device. Examples of such moieties include, but are not limited to, proteins, e.g., antibodies, antigens, enzymes, co-enzymes, e.g., biotin.

Herein, the term "punctuating endonuclease" refers to a restriction endonuclease which has the ability to cleave DNA at least one time. Typically, the punctuating enzyme recognizes a four base pair recognition sequence in a eukaryotic DNA. Preferably, the punctuating endonuclease cleaves DNA about every 256 base pairs. In a preferred embodiment, the punctuating endonuclease is the same as the first restriction endonuclease described herein. Examples of punctuating endonucleases useful in the methods of the present invention include, but are not limited to, Sau3A, MspI, MboI, AluI, BstUI, DpnII, HaeIII, HhaI, HinPI, MseI, NlaIII, RmaI, and TaqI.

Herein, the term "affinity capture device" refers to a moiety which interacts with (e.g., binds to) the affinity capture label. The affinity capture device can further include a solid support, e.g., an insoluble matrix, e.g., a magnetic bead, covalently coupled to the capture moiety. Examples of such moieties include proteins, e.g., antibodies, antigens, enzymes. When the affinity capture label is biotin, a preferred protein capture moiety is streptavidin.

Herein, the phrase "adjacent to" refers to the physical location of a nucleotide or amino acid sequence (or a portion thereof) relative to another nucleotide or amino acid sequence (or a portion thereof). Typically, a sequence is adjacent to another sequence if it is within about 8, 10, 12, 14 or 15 base pairs or amino acids of the other sequence.

Herein, the term "5' underhang" refers to a single-stranded sequence located at the 5' end of the bottom strand of an adapter.

Herein, the term "5' overhang" refers to a single-stranded sequence located at the 5' end of the top strand of an adapter.

Herein, the phrase "compatible with" means that at least a portion of a given sequence, e.g., an overhang or underhang sequence, is complementary to a selected sequence, e.g., another overhang or underhang sequence. For example, a 3' overhang sequence of a first DNA molecule is compatible with a 3' underhang sequence of a second DNA molecule. In the present disclosure, the term "complementary" has its usual meaning from molecular biology. Two nucleotide sequences or strands are complementary if they have sequences which would allow base pairing (Watson-Crick or Hoogstein) according to the usual pairing rules. This does not require that the strands would necessarily base pair at every nucleotide; two sequences can still be complementary with a low level (e.g., about 1 - 3%) of base mismatch such as that created by deletion, addition, or substitution of one or a few (e.g., up to 5 in a linear chain of 25 bases) nucleotides, or a combination of such changes.

The present invention also provides DNA vectors and kits for use in the TALESTB embodiment. A preferred DNA vector, e.g., the TALESTB vector described herein, includes an insertion site and lacks punctuating endonuclease restriction sites. A preferred kit for use in identifying gene expression patterns in an mRNA population includes a DNA vector, e.g., the TALESTB vector described herein, a primer comprising about 7 to 40 T residues, a punctuating endonuclease, a first adapter which includes a double-stranded oligodeoxynucleotide sequence including a second restriction endonuclease (e.g., a Type IIs restriction endonuclease), recognition sequence, a 5' overhang sequence compatible with a first vector insertion site and a 5' underhang sequence compatible with a punctuating endonuclease restriction site, and a second adapter which includes a double-stranded oligodeoxynucleotide sequence including a 3'

underhang sequence and a 5' underhang sequence compatible with a second vector insertion site.

An overview of the third or TALESTB embodiment of the present invention is presented in Figure 5 and described below. Although the overview of the third embodiment described herein provides a detailed description of the invention using particular restriction endonucleases, and a defined vector, it is well known to those of skill in the art that other restriction endonucleases can be selected and other methods of molecular biology, such as those described in Sambrook J. et al., "Molecular Cloning: A laboratory Manual", Second Ed. (Coldspring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989, Volume 1, Chapter 7), can be used to practice the present invention and this invention is not limited to the detailed examples presented herein.

In order to perform the third embodiment of the present invention, polyadenylated mRNA is isolated from the cell population of interest using standard procedures. The mRNA is then converted to cDNA using reverse transcriptase by priming the mRNA with an oligo dT sequence that has a biotin group as its 5' end. The first strand cDNA is converted to double stranded cDNA using RNAaseH and DNA Pol I, again using standard procedures. The double-stranded cDNA is then digested with a restriction enzyme, e.g., Sau3A. This enzyme has a 4-base recognition sequence (GATC) which occurs in eukaryotic DNA on average once every 256 base pairs and will cleave the average cDNA molecule several times. The 3' most fragment (representing the sequence between the 3' -most Sau3A site and the poly-A tail of each cDNA) is then captured by affinity capture on magnetic beads covalently coupled to streptavidin and all other Sau 3A restriction fragments are washed away leaving protruding fragments of the following partially double-stranded sequence (made up of SEQ ID NO:7 and SEQ ID NO:8, wherein N can be any of A, T, C or G):

GATCNNNNNNNNNN . . . NNNAAAAAAA . . . A
NNNNNNNNN . . . NNNTTTTTTT . . . T--Solid Phase

The next step is to anneal the solid phase (on magnetic bead) cDNA to a synthetic double stranded oligonucleotide first adapter having the partially double-stranded sequence (made up of SEQ ID NO:9 and SEQ ID NO:10):

5'- GGCCGCCGACTAGTGCAC-3'
3'-CGGCTGATCACGTC**CTAG**-5'

wherein the overhanging "CTAG" sequence on the lower strand will anneal to the overhanging "GATC" sequence on the solid phase cDNA molecules. This adapter sequence contains a BsgI restriction site (GTGCAG) which is located immediately 5' to the GATC sequence in the annealed cDNA. BsgI is a Type IIs restriction enzyme which recognizes the defined sequence shown above, but cleaves the DNA some 16 bases "downstream" (3') from the recognition sequence. Cleavage of the solid phase cDNA with BsgI releases a partially double-stranded oligomeric sequence (made up of SEQ ID NO:11 and SEQ ID NO:12) from the magnetic beads with a defined sequence consisting of the adapter molecule and an additional cDNA-derived antisense strand leaving a 2-base 3' "overhang" as shown below:

GGCCGCCGACTAGTGCAGGATCNNNNNNNNNNNN
CGGCTGATCACGTCCTAGNNNNNNNNNN

The 5' end of this oligomer contains an unpaired "GGCC" sequence which is compatible with a NotI restriction site. This fragment is then annealed and ligated in solution phase to a second partially double-stranded adapter sequence (made up of SEQ ID NO:13 and SEQ ID NO:14) consisting of 16 degenerate oligonucleotides of sequence:

5'- GATCAGTTTAAACAG-3'
3'-NNCTAGTCAAATTTGTCTTAA-5'

The presence of the degenerate "NN" sequence allows the annealing of this adapter to the first ligation product to generate a second partially double-stranded ligation product (made up of SEQ ID NO:15 and SEQ ID NO:16) as shown:

GGCCGCCGACTAGTGCAGGATCNNNNNNNNNNNN**GATC**AGTTTAAACAG
CGGCTGATCACGTC**CTAG**NNNNNNNNNNNNCTAGTCAAATTTGTCTTAA

This new fragment then consists of 12 bases of unknown sequence derived from each cDNA which is flanked on both sides by a Sau3A site (GATC) and ends that are compatible with vectors digested with NotI and EcoRI respectively. It will be understood by those of ordinary skill in the art that the sequences compatible with vector insertion sites on a first and second adapter are interchangeable, i.e., the first adapter can have a sequence compatible with a NotI insertion site and the second adapter can have a sequence compatible with an EcoRI insertion site as described above, or vice versa as

presented in Figure 5. When these inserts are cloned into such cut vectors a new cDNA "tag" library is formed in which each mRNA species generates a defined 12-base sequence. The library is cloned into the TALESTB plasmid vector and transformed into a suitable *E. coli* host. Plasmid DNA is then isolated by standard procedures and digested with Sau3A to release the partially double-sranded "tag" sequence (made up of SEQ ID NO:17 and SEQ ID NO:18)

GATCNNNNNNNNNNNN

NNNNNNNNNNNNCTAG

where the 12 "Ns" represent unknown sequence derived from the cDNA inserts. In order to separate these tags from other small restriction fragments derived from Sau3A sites within the plasmid backbone, certain of these sites were destroyed in the TALESTB vector by site-directed mutagenesis. The TALEST tags are isolated by gel electrophoresis, mixed together and subjected to enzymatic treatment with DNA ligase in order to generate tandem arrays of 50-60 tags in a single molecule. The arrays are then cloned into a sequencing vector and subjected to automated DNA sequence analysis.

When the arrays are analyzed, individual tags are recognized because they are separated from each other by the defined GATC punctuation sequence derived from the 3' -most Sau3A site in the original cDNA. Every tag has 12 additional bases of hitherto unknown sequence and uniquely marks one of the expressed genes in the cell population under study. Tags can ligate into the array in either sense or antisense orientation. However, with 12 bases of unknown sequence 4^{12} or 16,777,216 possible different tag sequences. This number exceeds by more than two orders of magnitude the number of expressed genes in the human genome (in all tissues). Hence, it is virtually impossible that a given tag sequence will match one gene in its sense orientation and a different gene in its antisense orientation. Moreover, the presence of the GATC start sequence effectively provides an additional 4 bases of identifying information and also localizes that information to a particular site within the tagged gene. However, in order to generate a frequency distribution of individual tags, it is important to consider tags in both sense and antisense orientation as identical. In order to accomplish this, a software program was produced. This software program scans automated DNA sequence files for pairs of restriction endonuclease sequences (e.g., punctuating restriction endonuclease sequences, e.g., GATC) interspersed with random sequence of defined length (e.g., 12 base pairs as generated when the TALEST embodiment is performed using the restriction endonuclease BsgI). The software then parses the sequence into individual

tags consisting of the base-pair segment between each pair of restriction endonuclease sequences. Multiple encounters of the same tag sequence are parsed together to generate a frequency distribution of tags. Because a tag can ligate into a tag array in either sense or antisense orientation, the software should establish a method to score tags as identical regardless of the orientation. This is accomplished by establishing the convention that every tag sequence identified by the software is compared with its reverse complement sequence, and only the sequence which is alphabetically primary is entered in the frequency distribution. The software can also compare frequency distributions of tags generated from different cells or tissues and highlight those whose frequency differs by any user designated level.

Automated high throughput DNA sequencers known to those of skill in the art allow simultaneous sequence determination of the tags. Thus, this method provides a simple and rapid way of producing tags that can be easily and quickly analyzed using high throughput DNA sequencers. Furthermore, because the present method involves the initial generation of a cDNA library, that library can be probed with an oligonucleotide corresponding to any tag of interest to determine the frequency of expression of the gene identified by the tag. For example, if a given tag shows up three times in a tumor cDNA pool but not at all in the normal cell pool, both cDNA libraries could be probed with a tag to ascertain their exact frequencies. A full length gene could then be isolated and identified using cloning methods known to those of skill in the art.

Another embodiment related to the TALESTB embodiment (diagramed in Figure 5), is presented schematically in Figure 6. In this embodiment, only a single adapter is used and the same steps as used in the TALESTB embodiment are used to isolate cDNA fragments that has been captured by the affinity capturing device. That is, a cDNA population is prepared from a preparation of mRNA using a primer covalently linked to an affinity capture label (e.g., biotin), and the cDNA is then cleaved with a punctuating restriction endonuclease (e.g., MboI or Sau3A) which cleaves only in the cDNA sequences. The 3' cDNA fragments are then captured using an affinity capture device (e.g., streptavidin linked to magnetic beads), and the uncaptured fragments are washed away. The captured cDNA inserts are then annealed and ligated to an adapter which is a double -stranded oligodeoxynucleotide sequence having an end compatible with the ends of the cDNA inserts (i.e., compatible with the punctuating restriction endonuclease site), a Type IIs restriction endonuclease recognition sequence (e.g., the recognition sequence of BsgI) and an end compatible with an EcoRI restriction site. The cDNA are then cleaved from the affinity capture device using the Type IIs restriction endonuclease (e.g., BsgI as shown in Figure 6) and the cDNA fragments are isolated.

At this point in the method, instead of providing a second adapter as shown in Figure 5, a vector having a restriction endonuclease acceptor site compatible with an end of the ligated adapter (e.g., an EcoRI site) and a site compatible with the other end of the cDNA molecules (i.e., an underhang sequence that can anneal with the restriction site of the BsgI enzyme) is provided. Preferably, to accept all of the possible cDNA ends generated by BsgI cutting of the cDNA , the vector is a 16-fold degenerate set of plasmid vectors having a 2-base degenerate 3' underhang shown as "NN" in Figure 6. The cDNA and the vector are annealed and ligated to produce constructs that are introduced into a suitable host cell (e.g., *E. coli*) and amplified using standard techniques well known to those skilled in the art The amplified plasmids are isolated and digested with the punctuating restriction endonuclease (e.g., Sau3A) to release the cDNA tag sequences which are then isolated and ligated to produce tag arrays, usually of at least 10 tags and preferably of about 40-60 tags per array. The tag arrays are then cloned using standard techniques into a suitable vector (e.g., a plasmid cut with BamHI to provide ends compatible with the punctuating restriction endonuclease sites) and the tag arrays are then sequenced. As shown in Figure 6, the nucleotide sequence of a tag array will consist of a punctuating restriction endonuclease sequence (GATC as shown in Figure 6), followed by a cDNA sequence, followed by another punctuating restriction endonuclease sequence, followed by a cDNA sequence, and so on until the flanking vector sequence. Thus, gene expression patterns in the mRNA population can be identified by identifying the tag sequences, each of which represents an expressed gene.

Other Embodiments

The three embodiments of the present invention are useful in the additional methods described below.

For example, the methods of the present invention can be used to determine the frequency of gene expression in an mRNA population. The method includes preparing the DNA constructs comprising the cDNA inserts of the invention, to obtain a cDNA library. The method further includes preparing an oligonucleotide probe comprising a tag sequence that is of interest, preferably using the methods of the present invention to identify a gene that is differentially expressed, and probing the cDNA library with an oligonucleotide probe comprising the tag sequence to determine the frequency of expression of a gene which includes the tag sequence.

The term "oligonucleotide probe" refers to a nucleic acid which specifically binds to a molecule of interest.

The term "probing" is used herein to refer to the method by which a nucleotide sequence, such as a nucleotide sequence comprising a tag, is used to hybridize to a pool

of RNA or DNA. The pool RNA or DNA can be isolated from its natural environment in the cell or tissue, or the pool can be assayed *in situ*, within the cell or tissue.

As used above and throughout this application, "hybridize" has its usual meaning from molecular biology. It refers to the formation of a base-paired interaction between nucleotide polymers. The presence of base pairing implies that a fraction of the nucleotides (e.g., at least 80%) in each of two nucleotide sequences are complementary to the other according to the usual base pairing rules. The exact fraction of the nucleotides which must be complementary in order to obtain stable hybridization will vary with a number of factors, including nucleotide sequence, salt concentration of the solution, temperature, and pH.

In referring to hybridization under "stringent conditions", "stringent" should be understood as an empirical term for any one nucleic acid sequence. However, the term indicates that the nature of the hybridization conditions is such that DNA sequences with an exact match for base pairing, or only a small percentage (5-10%) of base mismatch between the two sequences, will form base paired hybrid molecules which are stable enough to allow detection and isolation. On the other hand, two sequences with a higher level of base mismatch will not form such a stable hybrid under the same conditions. One skilled in the art will know that various factors can be altered to modulate the stringency of the conditions, and will understand how to alter those factors to obtain a desired effect. Examples of these factors are temperature, concentration of sodium ion, and concentration of tetramethylammonium chloride or tetraethylammonium chloride. One skilled in the art will recognize that the degree of stringency of a given set of conditions will be affected by characteristics of the DNA or RNA such as G+C content of the molecules, length of the shorter molecule, and location of the mismatches along the molecules. However, one skilled in the art will also know that there exist formulae which allow an estimation of the melting temperature (T_M). An example, for DNA, of such a formula for oligonucleotide probes is a function based on variables for sodium ion concentration, G+C content, and probe length. (Sambrook *et al., Molecular Cloning* (1989) at 11.46). Similar formulas are available for RNA:RNA hybrids and RNA:DNA hybrids. (*Id.* at 9.51.) In addition, one skilled in the art will know that the effect of mismatches on melting temperature can be estimated, and that melting temperature can be determined empirically for DNA sequences with perfect matching or with mismatches.

Therefore, one skilled in the art would recognize that "stringent conditions" can be readily determined for the claimed DNA sequences using only routine techniques. In this invention, "stringent conditions" should preferably require at least 80% base pairing,

more preferably at least 90% or 95% base pairing, still more preferably at least 97% base pairing, and most preferably at least 98% base pairing.

Those of skill in the art will recognize that the hybridization conditions can be varied by varying temperature, salt concentration, and formamide content of the hybridization and washing solutions. In addition, allowances can be made in the conditions for level of possible mismatch, or to provide a higher or lower level of stringency. Also, the proper level of stringency can be determined empirically to provide specific hybridization using the calculated T_M as a starting estimate. For example, the correspondence of T_m and the degree of mismatch may be calculated according to methods known to those skill in the art, as well as according to the methods described in, for example, Sambrook et al., Molecular Cloning (1989) at 11.47, 11.55-57.

The methods of the present invention can also be used to detect a difference in gene expression between two or more mRNA populations. The method includes identifying a gene expression pattern from a first mRNA population and from at least one additional mRNA population according to the methods of the present invention. The gene expression patterns so obtained can then be compared, thereby detecting differences in gene expression between the mRNA populations. In preferred aspects, the first mRNA population is obtained from a normal cell or tissue and the additional mRNA population is obtained from a cell or tissue from a target organism having a disease or disorder. In other preferred aspects, the mRNA populations are obtained from cells or tissues at different developmental stages. In yet other preferred aspects, the mRNA populations are obtained from cells derived from different tissues or organs of the same target organism. In other preferred aspects, the mRNA populations are obtained from different target organisms.

For purposes of the present invention, the term "target organism" includes any organism from which RNA can be obtained. Those skilled in the art will recognize that the term includes, for example, animals, plants, other eukaryotic cells, and bacteria.

The present invention also provides a method for detecting the presence of a disease in a target organism. The method includes identifying a gene that is expressed differently in a normal cell or tissue than in a cell or tissue from a target organism having a disease or disorder according to the methods of the present invention and isolating the tag sequence of the gene. An mRNA population obtained from a first target organism and an mRNA population obtained from a second normal or diseased target organism can be probed with the tag sequence to determine the level of expression of the gene. The level of expression of the gene in the first target organism can then be compared

with the level of expression of the gene in the second target organism to detect the presence of a disease in the first target organism.

In yet another embodiment, the methods of the invention can be used to isolate a gene. To isolate a gene, a tag that comprises a portion of the sequence of the gene to be isolated is identified and the gene is isolated by standard techniques, e.g., use of the tag sequence as a probe to probe to identify full length clones from a cDNA library. A "portion" of the sequence of the gene to be isolated refers to a linear chain that has a nucleotide sequence which is the same as a sequential subset of the sequence of the chain to which the portion refers.

In a preferred embodiment, the methods of the invention can be used to isolate a differentially expressed gene or a gene that is expressed at different levels in a first mRNA population compared to a second mRNA population. To isolate a differentially expressed gene, the nucleotide sequence of ligated tag arrays is obtained from a first cell type or tissue and from a second cell type or tissue according to the methods of the present invention. The frequency of expression of the individual tag sequences of the first and second cell types or tissues are then compared. Differentially expressed tag sequences in the first cell type or tissue compared to the second cell type or tissue can then be identified and isolated. A gene corresponding to the differentially expressed tag sequences can then be identified. By the term "correspond," is meant that at least a portion of one nucleic acid molecule is either complementary or homologous to a second nucleic acid molecule. Thus, a cDNA molecule may correspond to the mRNA molecule where the mRNA molecule was used as a template for reverse transcription to produce the cDNA molecule. Similarly, a genomic sequence of a gene may correspond to a cDNA sequence where portions of the genomic sequence are homologous or complementary to the cDNA sequence.

To isolate a gene which is expressed at different levels in a first mRNA population compared to a second mRNA population, a gene expression pattern from a first mRNA population and a gene expression pattern from a second additional mRNA population is identified according to the methods described herein. The gene expression patterns can then be compared to detect differences in gene expression between the mRNA populations. A gene that is expressed at a different level in the first mRNA population compared to the second mRNA population can then be identified and isolated.

This invention is further illustrated by the following examples which should not be construed as limiting. The contents of all references, patent applications, patents, and published patent applications cited throughout this application are hereby incorporated by reference.

EXAMPLES

The methods described in Examples 1, 2, 4, 5, and 6 can be used in each of the three embodiments of the methods described herein. Example 3 describes methods for generating tags in each of the three embodiments described herein.

EXAMPLE 1 - ISOLATION OF mRNA

Methods of extraction of RNA are well known in the art and are described, for example, in Sambrook J., et al., "Molecular Cloning: A Laboratory Manual", Second Ed. (Coldspring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989, Volume 1, Chapter 7). Other isolation extraction methods are also well known. Isolation is particularly performed in the presence of chaotropic agents such as guanadinium chloride or guanadinium isothiocyanate, other detergents and extraction agents can alternatively be used. It is desirable, but not required, that the messenger RNA be isolated from the total extract RNA by chromatography over an oligo (dT)-cellulose column or other, chromatographic media that have the capability of binding the polyadenylated 3' portion of the mRNA molecules.

Briefly, cells are lysed in RNA extraction buffer [0.14 M NaCl, 1.5 mM $MgCl_2$, 10 mM TrisHCl (pH 8.6), 0.5% NP-40, 1 mM DTT, 1000 units/ml RNase inhibitor (Pharmacia)] by using a Vortex mixer for 30 sec and then left standing on ice for 5 min. Nuclei and other cell debris were precipitated by centrifuging at 12,000 g for 90 sec, and the supernatant was deproteinized with Proteinase K followed by phenol extraction. RNA was precipitated by isopropanol and rinsed with 70% ethanol. Finally, the poly A+ fraction was collected by oligo dT column fractionation (Aviv, D. P., *et al., Proc. Natl. Acad. Sci.* USA 69, 1408-1412 (1972)).

EXAMPLE 2 - PREPARATION OF DOUBLE STRANDED cDNA

Double stranded cDNA is then prepared from the mRNA population using a DNA primer of the sequence depicted in Figure 3. The anchor primer includes a tract of T residues (approximately 7-40 T residues) and a site for cleavage by a restriction enzyme which recognizes more than 6 bases, the site for cleavage being located to the 5' site of the tract of T residues, such as NotI. The cDNA reaction is carried out under conditions that are well known in the art. Such techniques are described in, for example, Volume 2 of J. Sambrook et al., "Molecular Cloning: A Laboratory Manual., Second Ed.". In these methods, one way to carry out this method is by using reverse transcriptase from avion myeloblastosis virus.

The second cDNA strand synthesis may be performed using the RNAase H/DNA polymerase I self priming method. Briefly, two micrograms each of the cytoplasmic Poly A^+ RNA and the vector primer DNA were co-precipitated in 70% ethanol containing 0.3 M Na-acetate and the pellet was dissolved in 12 Fl of distilled water. For the first strand synthesis, after heat denaturation at 76°C for 10 min, 4 Fl of 5X reaction buffer (250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$), 2 Fl of 0.1 M DTT, 1 Fl of 10 mM each of dATP, dCTP, dGTP and dTTP were added to the sample at 37°C. The reaction was initiated by the addition of 200 units of reverse transcriptase MMLV-H-RT (BRL), and after incubation at 37°C for 30 min, stopped by transferring the reaction tube onto ice. For the second strand synthesis, to the aforementioned reaction mixture were added 92 Fl of distilled water, 32 Fl of 5X *E. coli* reaction buffer (100 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$,, 50 mM $(NH4)_2SO_4$, 500 mM KCl, 250 g/ml of BSA, 750 M ßNAD), 3 Fl of 10 mM each of dATP, dCTP, dGTP and dTTP, 15 units of *E. coli* ligase (Pharmacia), 40 units of *E. coli* polymerase (Pharmacia), and 15 units of RNase H (Pharmacia), which was then incubated at 16°C for 2 h. The reaction mixture was heated to 65°C for 15 min.

The cDNA sample is then cleaved with MboI and NotI. The cDNA vector sample is then inserted into the TALEST vector depicted in Figure 2. The TALEST vector has similarly been digested with Bam H1 and NotI using methods known to those skilled in the art. Briefly, a sample containing blank cDNA inserts and blank vector is diluted to up to one ml with 1x *E. coli* reaction buffer, and 100 units of *E. coli* ligase are added. The resulting mixture is incubated at 16 "C overnight. Following insertion of the cDNA, the vector mixture is then used to transform *E. coli* competent cells. Suitable host cells for cloning are described in, for example, Sambrook et al., "Molecular Cloning: A Laboratory Manual". The host cell is grown to increase or amplify the number of vectors produced. A suitable *E. coli* strain is DH5 or MC1061.

EXAMPLE 3 - GENERATION OF TAGS

In the TALEST embodiment, the TALEST vectors are isolated from the grown host cell using methods known by those skilled in the art, such as those described for "minipreps," described in, for example, J. Sambrook et al., "Molecular Cloning: A Laboratory Manual., Second Ed." The vectors are then cleaved with BsgI which linearizes the plasmid at a site 12 bases downstream from the MboI start sequence on the sense strand and 10 bases on the antisense strand. T4 DNA polymerase is then used to generate blunt ends on the vector. The vectors are then cleaved with PmlI which results in a 20 base blunt ended fragment with the sequence GTGCAGGATCNNNNNNNNNN.

The tags are separated from the remainder of the vector using polyacrylamide gel electrophoresis as described in , for example, Sambrook et al., *supra*.

In the TALESTA embodiment, the double stranded cDNA is cleaved with the restriction endonuclease Sau3A to generate restriction fragments, and the 3' most fragment containing the oligodT-bioting moiety is captured using streptavidin magnetic beads. The fragment has the partially double-strabded sequence (made up of SEQ ID NO:7 and SEQ ID NO:8) as shown:

```
GATCNNNNNNNNN. . .NNNAAAAAAA. . .A
    NNNNNNNNN. . .NNNTTTTTTT. . .T-Biotin -
```

The captured fragment, still affixed to the magnetic bead, is then annealed to a 5' adapter having the partially double-stranded sequence (made up of SEQ ID NO:19 and SEQ ID NO:20):

```
AATTCGACTAGTGCAG
    GCTGATCACGTCCTAG
```

to generate a ligated complex having the double-stranded sequence (made up of SEQ ID NO: 21 and SEQ ID NO:22):

```
AATTCGACTAGTGCAGGATCNNNNNNNNN. . .NNNAAAAAAA. . .A
    GCTGATCACGTCCTAGNNNNNNNNN. . .NNNTTTTTTT. . .T-Biotin -
```

Digestion of the solid-phase bound cDNA with the Type IIs restriction endonucelase BsgI cleaves the cDNA insert at a defined distance from the 5' end releasing a fragment having the partially double-stranded sequence (made up of SEQ ID NO:23 and SEQ ID NO:24):

```
AATTCGACTAGTGCAGGATCNNNNNNNNNNNN
    GCTGATCACGTCCTAGNNNNNNNNN
```

The released fragment is then cloned into a 16-fold degenerate vector into a cloning site having the sequence:

```
...G                                GATC...
...CTTAA                            NNCTAG...
```

The fragment is ligated into the vector, transformed into competent *E. coli* and plasmid DNA is prepared. Plasmid DNA is then digested with the restriction endonuclease Sau3A to release the tag having the partially double-stranded sequence (made up of SEQ ID NO:17 and SEQ ID NO:18):

```
GATCNNNNNNNNNNNN
    NNNNNNNNNNNNCTAG
```

In the TALESTB embodiment, double stranded cDNA is cleaved with the restriction endonuclease Sau3A to generate restriction fragments, and the 3' most fragment containing the oligodT-biotin moiety is captured using streptavidin magnetic beads. The fragment has the partially double-stranded sequence (made up of SEQ ID NO:7 and SEQ ID NO:8):

```
GATCNNNNNNNNN. . .NNNAAAAAAA. . .A
    NNNNNNNNN. . .NNNTTTTTTT. . .T-Biotin -
```

The captured fragment, still affixed to the magnetic bead, is then annealed to a 5' adapter having the partially double-stranded sequence (made up of SEQ ID NO:19 and SEQ ID NO:20):

```
AATTCGACTAGTGCAG
    GCTGATCACGTCCTAG
```

to generate a ligated complex having the partially double-stranded sequence (made up of SEQ ID NO:21 and SEQ ID NO:22):

```
AATTCGACTAGTGCAGGATCNNNNNNNNN. . .NNNAAAAAAA. . .A
    GCTGATCACGTCCTAGNNNNNNNNN. . .NNNTTTTTTT. . .T-Biotin -
```

Digestion of the solid-phase bound cDNA with the Type IIs restriction endonuclease BsgI cleaves the cDNA inserts at a defined distance from the 5' end releasing a fragment having the partially double starnded sequence (made up of SEQ ID NO:23 and SEQ ID NO:24):

```
AATTCGACTAGTGCAGGATCNNNNNNNNNNNN
    GCTGATCACGTCCTAGNNNNNNNNNN
```

The released fragment is then ligated to a 16-fold degenerate second adapter having the partially double-stranded sequence (made up of SEQ ID NO:13 and SEQ ID NO:14):

```
  GATCAGTTTAAACAGC
NNCTAGTCAAATTTGTCGCCGG
```

to yield a ligated fragment having the partially double-stranded sequence (made up of SEQ ID NO:25 and SEQ ID NO:26):

```
AATTCGACTAGTGCAGGATCNNNNNNNNNNNNNGATCAGTTTAAACAGC
     GCTGATCACGTCCTAGNNNNNNNNNNNNNCTAGTCAAATTTGTCGCCGG
```

The fragment is then ligated into a cloning vector which has been digested with the restriction endonucleases EcoRI and NotI to generate the following cloning site:

```
...G                                        GGCC...
...CTTAA                                       ...
```

The resultant recombinant vector is then transformed into competent E. coli and plasmid DNA is prepared. The plasmid DNA is then digested with the restriction endonuclease Sau3A to release the tag having the partially double-stranded sequence (made up of SEQ ID NO:17 and SEQ ID NO:18):

```
GATCNNNNNNNNNNNN
    NNNNNNNNNNNNCTAG
```

EXAMPLE 4 - SEQUENCING OF TAGS

The tags generated in Example 3 are mixed together and subjected to enzymatic treatment with DNA ligase in order to generate tandem arrays of 30-40 tags in a single molecule. To isolate lengths of 30-40 tags, DNA sequences of approximately 420-560 nucleotides in length are isolated by agarose gel electrophoresis as described in, for example, Sambrook et al., *supra.* The arrays of 30-40 tags are then cloned into a sequencing vector. Suitable sequencing vectors are known to those of skill in the art. One example of an appropriate sequencing vector is pUC19. The sequencing vector containing the tags is then subjected to automated DNA sequence analysis.

EXAMPLE 5 - DETERMINATION OF FREQUENCY OF GENE EXPRESSION BY PROBING CDNA LIBRARIES WITH TAG SEQUENCES

If a particular sequence tag appears to be over or under represented in any individual collection of tags, the actual frequency of the gene from which the tag was isolated may be determined by probing the parent cDNA library. Standard methods known to those skilled in the art may be used to probe the parent cDNA library. For example, prior to isolation of bacterial colonies for plasmid isolation and tag generation, the plates containing the colonies can be overlaid with a nitrocellulose or nylon membrane to generate a replica copy. Alternatively, a new cDNA library from the same tissue source can be produced in either plasmid or phage vectors and expose to filters as described above. The filters are then exposed to a synthetic oligonucleotide probe having the same sequence as the tag of interest. The probe is first labeled with ^{32}P using standard techniques as described in J. Sambrook et al., "Molecular Cloning: A Laboratory Manual.; Second Ed. and other sources. Filters are then washed and exposed to X-ray film. By counting the number colonies or plaques which hybridize to the probe and dividing that number by the total number of clones in the screened library, one obtains a frequency estimate of the transcript prevalence in the tissue from which the library was derived.

EXAMPLE 6 - CLONING OF DIFFERENTIALLY EXPRESSED GENES

The methods of the present invention may be used to isolate differentially-expressed genes. Particular relatively over-expressed genes may be identified and isolated. By comparing tag frequencies in different libraries derived from related tissues (for example, a tumor and the normal tissue from which it arose) it is possible to identify tags corresponding to genes that are over- or under- expressed in one of the tissues and may be responsible for a pathological or other phenotype of either tissue. In order to more fully characterize these "differentially expressed" genes, one can search the tag sequence against an appropriately filtered database of human RNA or cDNA sequences. Alternatively one can use the tag sequence as a hybridization probe as described in Example 5 to identify full-length clones from a cDNA library. These clones can then be sequenced and searched for homologies to known genes using standard procedures.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:
(A) NAME: Chugai Biopharmaceuticals, Inc.
(B) STREET: 6275 Nancy Ridge Drive
(C) CITY: San Diego
(D) STATE: California
(E) COUNTRY: USA
(F) POSTAL CODE (ZIP): 92121-4362

(ii) TITLE OF INVENTION: METHOD FOR ANALYZING QUANTITATIVE EXPRESSION OF GENES

(iii) NUMBER OF SEQUENCES: 26

(iv) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: ASCII text

(v) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: US 08/784,208
(B) FILING DATE: 15-JAN-1997

(viii) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: LAHIVE & COCKFIELD, LLP
(B) STREET: 28 STATE STREET
(C) CITY: BOSTON
(D) STATE: MASSACHUSETTS
(E) COUNTRY: USA
(F) ZIP: 02109

(ix) ATTORNEY/AGENT INFORMATION:
(A) NAME: Jean M. Silveri
(B) REGISTRATION NUMBER: 39,030
(C) REFERENCE/DOCKET NUMBER: GNI-004CPPC

(x) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: (617)227-7400
(B) TELEFAX: (617)742-4214

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 3737 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TCGCGCGTTT CGGTGATGAC GGTGAAAACC TCTGACACAT GCAGCTCCCG GAGACGGTCA 60

CAGCTTGTCT GTAAGCGGAT GCCGGGAGCA GACAAGCCCG TCAGGGCGCG TCAGCGGGTG 120

TTGGCGGGTG TCGGGGCTGG CTTAACTATG CGGCATCAGA GCAGATTGTA CTGAGAGTGC 180

ACCATATGCG GTGTGAAATA CCGCACAGAT GCGTAAGGAG AAAATACCGC ATCAGGCGCC 240

ATTCGCCATT CAGGCTGCGC AACTGTTGGG AAGGGCGATC GGTGCGGGCC TCTTCGCTAT 300

TACGCCAGCT GGCGAAAGGG GGATGTGCTG CAAGGCGATT AAGTTGGGTA ACGCCAGGGT 360

TTTCCCAGTC ACGACGTTGT AAAACGACGG CCAGTGAATT CGAGCTCGGT ACCGGATGAC 420

ACGTGCAGGA TCCATGATCA TCGTGGCGCA TGTATTACTC ATCCTTTTGG GGGCCACTGA 480

GATACTGCAA GCTGACTTAC TTCCTGATGA AAAGATTTCA CTTCTCCCAC CTGTCAATTT 540

CACCATTAAA GTTACTGGTT TGGCTCAAGT TCTTTTACAA TGGAAACCAA ATCCTGATCA 600

AGAGCAAAGG AATGTTAATC TAGAATATCA AGTGAAAATA AACGCTCCAA AAGAAGATGA 660

CTATGAAACC AGAATCACTG AAAGCAAATG TGTAACCATC CTCCACAAAG GCTTTTCAGC 720

AAGTGTGCGG ACCATCCTGC AGAACGACCA CTCACTACTG GCCAGCAGCT GGGCTTCTGC 780

TGAACTTCAT GCCCCACCAG GGTCTCCTGG AACCTCAATT GTGAATTTAA CTTGCACCAC 840

AAACACTACA GAAGACAATT ATTCACGTTT AAGGTCATAC CAAGTTTCCC TTCACTGCAC 900

CTGGCTTGTT GGCACAGATG CCCCTGAGGA CACGCAGTAT TTTCTCTACT ATAGGTATGG 960

CTCTTGGACT GAAGAATGCC AAGAATACAG CAAAGACACA CTGGGGAGAA ATATCGCATG 1020

CTGGTTTCCC AGGACTTTTA TCCTCAGCAA AGGGCGTGAC TGGCTTTCGG TGCTTGTTAA 1080

CGGCTCCAGC AAGCACTCTG CTATCAGGCC CTTTGATCAG CTGTTTGCCC TTCACGCCAT 1140

TGATCAAATA AATCCTCCAC TGAATGTCAC AGCAGAGATT GAAGGAACTC GTCTCTCTAT 1200

CCAATGGGAG AAACCAGTGT CTGCTTTTCC AATCCATTGC TTTGATTATG AAGTAAAAAT 1260

ACACAATACA AGGAATGGAT ATTTGCAGAT AGAAAAATTG ATGACCAATG CATTCATCTC 1320

AATAATTGAT GATCTTTCTA AGTACGATGT TCAAGTGAGA GCAGCAGTGA GCTCCATGTG 1380

CAGAGAGGCA GGGCTCTGGA GTGAGTGGAG CCAACCTATT TATGTGGGAA ATGATGAACA 1440

CAAGCCCTTG AGAGAGTGGT TTGTCGCGGC CGCTCTAGAG TCGACCTGCA GGCATGCAAG 1500

CTTGGCGTAA TCATGGTCAT AGCTGTTTCC TGTGTGAAAT TGTTATCCGC TCACAATTCC 1560

ACACAACATA CGAGCCGGAA GCATAAAGTG TAAAGCCTGG GGTGCCTAAT GAGTGAGCTA 1620

ACTCACATTA ATTGCGTTGC GCTCACTGCC CGCTTTCCAG TCGGGAAACC TGTCGTGCCA 1680

GCTGCATTAA TGAATCGGCC AACGCGCGGG GAGAGGCGGT TTGCGTATTG GGCGCTCTTC 1740

CGCTTCCTCG CTCACTGACT CGCTGCGCTC GGTCGTTCGG CTGCGGCGAG CGGTATCAGC 1800

TCACTCAAAG GCGGTAATAC GGTTATCCAC AGAATCAGGG GATAACGCAG GAAAGAACAT 1860

GTGAGCAAAA GGCCAGCAAA AGGCCAGGAA CCGTAAAAAG GCCGCGTTGC TGGCGTTTTT 1920

CCATAGGCTC CGCCCCCCTG ACGAGCATCA CAAAAATCGA CGCTCAAGTC AGAGGTGGCG 1980

AAACCCGACA GGACTATAAA GATACCAGGC GTTTCCCCCT GGAAGCTCCC TCGTGCGCTC 2040

TCCTGTTCCG ACCCTGCCGC TTACCGGATA CCTGTCCGCC TTTCTCCCTT CGGGAAGCGT 2100

GGCGCTTTCT CATAGCTCAC GCTGTAGGTA TCTCAGTTCG GTGTAGGTCG TTCGCTCCAA 2160

GCTGGGCTGT GTGCACGAAC CCCCCGTTCA GCCCGACCGC TGCGCCTTAT CCGGTAACTA 2220

TCGTCTTGAG TCCAACCCGG TAAGACACGA CTTATCGCCA CTGGCAGCAG CCACTGGTAA 2280

CAGGATTAGC AGAGCGAGGT ATGTAGGCGG TGCTACAGAG TTCTTGAAGT GGTGGCCTAA 2340

CTACGGCTAC ACTAGAAGGA CAGTATTTGG TATCTGCGCT CTGCTGAAGC CAGTTACCTT 2400

CGGAAAAAGA GTTGGTAGCT CTTGATCCGG CAAACAAACC ACCGCTGGTA GCGGTGGTTT 2460

TTTTGTTTGC AAGCAGCAGA TTACGCGCAG AAAAAAAGGA TCTCAAGAAG ATCCTTTGAT 2520

CTTTTCTACG GGGTCTGACG CTCAGTGGAA CGAAAACTCA CGTTAAGGGA TTTTGGTCAT 2580

GAGATTATCA AAAAGGATCT TCACCTAGAT CCTTTTAAAT TAAAAATGAA GTTTTAAATC 2640

AATCTAAAGT ATATATGAGT AAACTTGGTC TGACAGTTAC CAATGCTTAA TCAGTGAGGC 2700

ACCTATCTCA GCGATCTGTC TATTTCGTTC ATCCATAGTT GCCTGACTCC CCGTCGTGTA 2760

GATAACTACG ATACGGGAGG GCTTACCATC TGGCCCCAGT GCTGCAATGA TACCGCGAGA 2820

CCCACGCTCA CCGGCTCCAG ATTTATCAGC AATAAACCAG CCAGCCGGAA GGGCCGAGCG 2880

CAGAAGTGGT CCTGCAACTT TATCCGCCTC CATCCAGTCT ATTAATTGTT GCCGGGAAGC 2940

TAGAGTAAGT AGTTCGCCAG TTAATAGTTT GCGCAACGTT GTTGCCATTG CTACAGGCAT 3000

CGTGGTGTCA CGCTCGTCGT TTGGTATGGC TTCATTCAGC TCCGGTTCCC AACGATCAAG 3060

GCGAGTTACA TGATCCCCCA TGTTGTGCAA AAAAGCGGTT AGCTCCTTCG GTCCTCCGAT 3120

CGTTGTCAGA AGTAAGTTGG CCGCAGTGTT ATCACTCATG GTTATGGCAG CACTGCATAA 3180

TTCTCTTACT GTCATGCCAT CCGTAAGATG CTTTTCTGTG ACTGGTGAGT ACTCAACCAA 3240

GTCATTCTGA GAATAGTGTA TGCGGCGACC GAGTTGCTCT TGCCCGGCGT CAATACGGGA 3300

TAATACCGCG CCACATAGCA GAACTTTAAA AGTGCTCATC ATTGGAAAAC GTTCTTCGGG 3360

GCGAAAACTC TCAAGGATCT TACCGCTGTT GAGATCCAGT TCGATGTAAC CCACTCGTGC 3420

ACCCAACTGA TCTTCAGCAT CTTTTACTTT CACCAGCGTT TCTGGGTGAG CAAAAACAGG 3480

AAGGCAAAAT GCCGCAAAAA AGGGAATAAG GGCGACACGG AAATGTTGAA TACTCATACT 3540

CTTCCTTTTT CAATATTATT GAAGCATTTA TCAGGGTTAT TGTCTCATGA GCGGATACAT 3600

ATTTGAATGT ATTTAGAAAA ATAAACAAAT AGGGGTTCCG CGCACATTTC CCCGAAAAGT 3660

GCCACCTGAC GTCTAAGAAA CCATTATTAT CATGACATTA ACCTATAAAA ATAGGCGTAT 3720

CACGAGGCCC TTTCGTC 3737

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 670 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(B) LOCATION: 6-304
(D) OTHER INFORMATION: /note= "N may be present or absent."

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(B) LOCATION: 368-666
(D) OTHER INFORMATION: /note= "N may be present or absent."

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GATCNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 60

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 120

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 180

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 240

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 300

NNNNAAAAAA AAAAAAAAAA AAAGCGGCCG CCATGCATGG CGGCCGCTTT TTTTTTTTTT 360

TTTTTTNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 420

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 480

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 540

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 600

NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN 660

NNNNNNGATC 670

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 32 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

TTTTTTTTTT TTTTTTTTTC GCCGGGCGCA TG 32

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GGATCNNNNN NNNNN 15

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

GTGCAGGATC NNNNNNNNNN 20

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 14 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GATCNNNNNN NNNN 14

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 23 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N can be 12 or more nucleic acid bases"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "A can be 7 or more A's"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GATCNNNNNN NNNNNNAAAA AAA 23

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 19 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N can be 12 or more nucleic acid bases"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "T can be 7 or more T's"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

NNNNNNNNNN NNTTTTTTT 19

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GGCCGCCGAC TAGTGCAC 18

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CGGCTGATCA CGTCCTAG 18

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 34 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GGCCGCCGAC TAGTGCAGGA TCNNNNNNNN NNNN 34

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CGGCTGATCA CGTCCTAGNN NNNNNNNN 28

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GATCAGTTTA AACAG 15

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

NNCTAGTCAA ATTTGTCTTA A 21

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 49 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GGCCGCCGAC TAGTGCAGGA TCNNNNNNNN NNNNGATCAG TTTAAACAG 49

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 49 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

CGGCTGATCA CGTCCTAGNN NNNNNNNNNN CTAGTCAAAT TTGTCTTAA 49

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GATCNNNNNN NNNNNN 16

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

NNNNNNNNNN NNCTAG 16

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

AATTCGACTA GTGCAG 16

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

GCTGATCACG TCCTAG 16

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 39 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N can be 12 or more nucleic acid bases"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "A can be 7 or more A's"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

AATTCGACTA GTGCAGGATC NNNNNNNNNN NNAAAAAAA 39

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 35 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N can be 12 or more nucleic acid bases"

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "T can be 7 or more T's"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

GCTGATCACG TCCTAGNNNN NNNNNNNNTT TTTTT 35

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 32 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

AATTCGACTA GTGCAGGATC NNNNNNNNNN NN 32

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GCTGATCACG TCCTAGNNNN NNNNNN 26

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 48 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

AATTCGACTA GTGCAGGATC NNNNNNNNNN NNGATCAGTT TAAACAGC 48

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 48 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(D) OTHER INFORMATION: /note= "N stands for A,C,T or G"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

GCTGATCACG TCCTAGNNNN NNNNNNNNCT AGTCAAATTT GTCGCCGG 48

What is claimed is:

1. A method for identifying gene expression patterns in a population of mRNA, comprising the steps of:

(a) preparing a population of double-stranded cDNA from a population of mRNA using a primer;

(b) cleaving said double-stranded cDNA with a first restriction endonuclease which cleaves at a site within said cDNA and not within said primer, to obtain a population of cDNA inserts;

(c) inserting said cDNA inserts into insertion sites of cloning vectors to obtain DNA constructs, wherein each cloning vector comprises a second restriction endonuclease recognition sequence located 5' to said insertion site, and a third restriction endonuclease recognition sequence located 5' to or overlapping with said second endonuclease recognition sequence;

(d) amplifying said DNA constructs in a host cell;

(e) isolating amplified DNA constructs;

(f) digesting said amplified DNA constructs with a second restriction endonuclease such that digestion of said DNA constructs with said second restriction endonuclease cleaves said DNA constructs at sites within said cDNA inserts;

(g) digesting said amplified DNA constructs with a third restriction endonuclease to obtain tags; and

(h) obtaining a nucleotide sequence of said tags to identify gene expression patterns in said population of mRNA.

2. The method of claim 1, wherein the obtaining step comprises:

ligating said tags to obtain a ligated tag array comprising at least about 10 tags;

inserting said ligated tag array into a vector; and

sequencing said ligated tag array.

3. The method of claim 1, wherein said first restriction endonuclease recognizes a sequence of four bases; wherein said second restriction endonuclease is a Type IIs restriction endonuclease; and wherein said third restriction endonuclease recognition sequence is located about 10 to 40 nucleotides 5' of said second restriction endonuclease cleavage site.

4. The method of claim 1, wherein said first restriction endonuclease recognizes a sequence of four bases; wherein said second restriction endonuclease is a Type IIs restriction endonuclease; and wherein said third restriction endonuclease recognition sequence overlaps said second restriction endonuclease recognition sequence.

5. The method of claim 1, wherein step (a) uses a primer comprising a priming restriction endonuclease cleavage sequence linked to a 5' end of an oligo dT sequence, and further comprising the step of digesting said double-stranded cDNA with a priming restriction endonuclease to obtain cDNA inserts comprising said priming restriction endonuclease cleavage sequence introduced at a 3' end of said double-stranded cDNA when said cDNA is digested with said priming restriction endonuclease.

6. The method of claim 2, wherein said ligated tag array comprises at least about 40 tags.

7. A method for identifying gene expression patterns in a population of mRNA, comprising the steps of:

(a) preparing a population of double-stranded cDNA from a first population of mRNA obtained from a first biological sample, using a primer covalently linked to an affinity capture label;

(b) cleaving said double-stranded cDNA with a punctuating restriction endonuclease which cleaves at a site within said cDNA and not within said primer, to obtain a population of cDNA inserts linked to said affinity capture label;

(c) capturing said cDNA inserts by capturing said affinity capture label with an affinity capture device to obtain a population of captured cDNA inserts;

(d) annealing a captured cDNA insert to a first adapter and ligating said cDNA insert and said first adapter to obtain a first ligation product, wherein said first adapter comprises a double-stranded oligodeoxynucleotide sequence comprising a 5' overhang sequence compatible with a first vector insertion site, a second restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a punctuating restriction endonuclease site;

(e) cleaving said first ligation product with a second restriction endonuclease to produce a released ligation product separated from said affinity capture label, wherein said released ligation product comprises a punctuating endonuclease restriction site adjacent to a cDNA sequence and a 3' overhang sequence;

(f) annealing said released ligation product with a second adapter and ligating said released ligation product and said second adapter to obtain a second ligation product, wherein said second adapter comprises a double-stranded oligodeoxynucleotide sequence comprising a 5' underhang sequence compatible with a second vector insertion site and a 3' underhang sequence compatible with said 3' overhang sequence of said released ligation product, and wherein said second ligation product comprises a 5' sequence compatible with a first vector insertion site, cDNA sequence flanked by punctuating endonuclease restriction sites, and a 3' sequence compatible with a second vector insertion site;

(g) inserting said second ligation product into a cloning vector at a first vector insertion site and a second vector insertion site to obtain a DNA construct;

(h) amplifying said DNA construct in a host cell;

(i) isolating amplified DNA constructs;

(j) digesting said amplified DNA constructs with said punctuating restriction endonuclease to obtain tags; and

(k) obtaining a nucleotide sequence of said tags to identify gene expression in said first biological sample.

8. The method of claim 7, wherein step (k) comprises:

ligating said tags to obtain a ligated tag array comprising at least about 10 tags, wherein each tag in said tag array is adjacent to a punctuating restriction endonuclease recognition site;

inserting said ligated tag array into a vector;

sequencing said ligated tag array; and

comparing sequences of said tag array to known gene sequences.

9. The method of claim 7, further comprising the step of isolating a gene sequence that hybridizes to a tag.

10. The method of claim 7, wherein step (a) uses an affinity capture label comprising biotin, and step (c) uses an affinity capture device comprising a magnetic bead covalently linked to streptavidin.

11. The method of claim 7, wherein step (e) uses a second restriction endonuclease that cleaves said first ligation product site at a site located about 16 nucleotides 3' of its recognition sequence.

12. The method of claim 7, wherein step (d) uses said first adapter comprising said second restriction endonuclease recognition site located 5' to sequence which is compatible with said punctuating restriction endonuclease site.

13. The method of claim 7, wherein step (e) produces said released ligation product comprising a 3' overhang of two nucleotides in length, and wherein step (f) uses said second adapter comprising a 3' underhang sequence comprising two nucleotides of degenerate sequence.

14. The method of claim 8, wherein said ligating step produces a ligated tag array of at least about 40 tags.

15. The method of claim 7, wherein step (e) cleaves said first ligation product with a second restriction endonuclease that is a Type IIs restriction endonuclease.

16. The method of claim 8, wherein step (a) uses said primer comprising a 5' oligo dT sequence covalently linked at a 3' end to a biotin label; wherein step (b) cleaves with Sau3A; wherein step (c) uses said affinity capture device comprising a magnetic bead covalently linked to streptavidin; wherein step (d) uses said first adapter comprising a 5' overhang sequence compatible with a NotI insertion site, a BsgI restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a Sau3A restriction site; wherein step (e) cleaves said first ligation product with BsgI to produce a released ligation product comprising a Sau3A restriction site adjacent to cDNA sequence; wherein step (f) uses said second adapter comprising a 5' underhang sequence compatible with an EcoRI insertion site and a 3' underhang degenerate sequence; wherein step (f) produces said second ligation product comprising a NotI insertion site, a cDNA sequence flanked by Sau3A restriction sites, and a EcoRI insertion site; wherein step (g) inserts said second ligation product into NotI and EcoRI sites of said cloning vector; wherein step (j) digests said amplified DNA constructs with Sau3A to obtain tags; and wherein said ligating step obtains ligated tag arrays of about 30 to 60 tags.

17. The method of claim 7, further comprising the steps of:

preparing an oligonucleotide probe comprising a nucleotide sequence of a tag; and

probing a cDNA library with said olignucleotide probe to determine a frequency of expression of a gene which comprises said tag.

18. The method of claim 7, further comprising the steps of:

repeating steps (a) through (k) using a second population of mRNA from a second biological sample; and

comparing gene expression of said first population of mRNA with gene expression of said second population of mRNA to determine differences in gene expression between said first biological sample and said second biological sample.

19. The method of claim 18, further comprising the steps of :

identifying a gene that is expressed at a first level in said first population of mRNA and is expressed at a second level in said second population of mRNA; and

isolating said gene from a cDNA library.

20. The method of claim 18 or 19, wherein said first biological sample is cells or tissue obtained from a normal non-diseased organism, and said second biological sample is cells or tissue obtained from an organism having a disease or disorder.

21. The method of claim 18 or 19, wherein said first biological sample is cells or tissue obtained from an organism at a first stage of development, and said second biological sample is cells or tissue obtained from an organism at a second stage of development.

22. A kit for identifying gene expression patterns in a population of mRNA according to the method of claim 7, comprising:

a DNA vector comprising a NotI insertion site, an EcoRI insertion site, and one or fewer Sau3A restriction endonuclease recognition sites;

a primer comprising about 7 to about 40 T residues;

a first adapter comprising a double-stranded oligonucleotide sequence comprising a 5' overhang sequence compatible with a NotI insertion site, a Type IIs restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a Sau3A restriction site; and

a second adapter comprising a double-stranded oligonucleotide sequence comprising a 5' underhang sequence compatible with an EcoRI insertion site and a 3' underhang degenerate sequence.

23. A method for identifying gene expression patterns in a population of mRNA, comprising the steps of:

a) preparing a population of double-stranded cDNA from a population of mRNA obtained from a biological sample, using a primer covalently linked to an affinity capture label;

b) cleaving said double-stranded cDNA with a punctuating restriction endonuclease which cleaves at a site within said cDNA and not within said primer, to obtain a population of cDNA inserts linked to said affinity capture label;

c) capturing said cDNA inserts by capturing said affinity capture label with an affinity capture device to obtain a population of captured cDNA inserts;

d) annealing a captured cDNA insert to an adapter and ligating said cDNA insert and said adapter to obtain a first ligation product, wherein said adapter comprises a double-stranded oligodeoxynucleotide sequence comprising a 5' overhang sequence compatible with a first vector insertion site, a Type IIs restriction endonuclease recognition sequence, and a 5' underhang sequence compatible with a punctuating restriction endonuclease site;

e) cleaving said first ligation product with a Type IIs restriction endonuclease to produce a released ligation product separated from said affinity capture label, wherein said released ligation product comprises a punctuating endonuclease restriction site adjacent to a cDNA sequence and a 3' overhang sequence of 2 nucleotides;

f) providing a vector comprising a restriction endonuclease acceptor site compatible with an end of the ligated adapter and a 3' underhang sequence of 2 degenerate nucleotides;

g) annealing said vector of step f) with said released ligation product of step e) to produce DNA constructs;

h) amplifying said DNA constructs in a host cell;

i) isolating said DNA constructs from said host cell and digesting said isolated DNA constructs said punctuating restriction endonuclease to release cDNA tag sequences;

j) isolating and ligating released cDNA tag sequences to produce tag arrays;

k) cloning tag arrays into a vector for DNA sequencing.

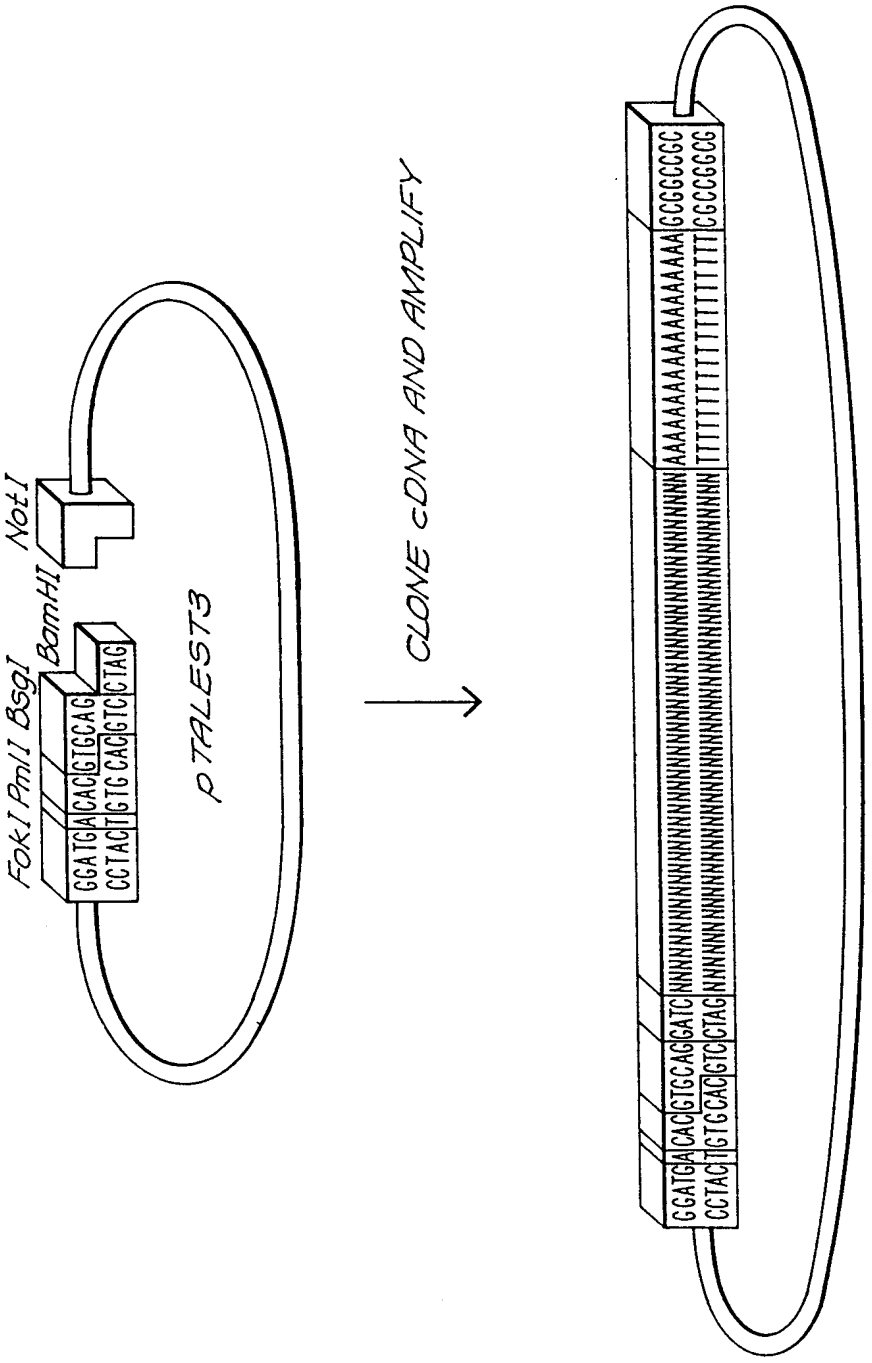

FokI PmlI BglI
BamHI
NotI
GGATGACACGTGCAG
CCTACTGTGCACGTC
CTAG
pTALEST3
CLONE cDNA AND AMPLIFY
GGATGACACGTGCAGGATC
CCTACTGTGCACGTCCTAG
AAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTT
GCGGCCGC
CGCCGGCG

FIG. 1

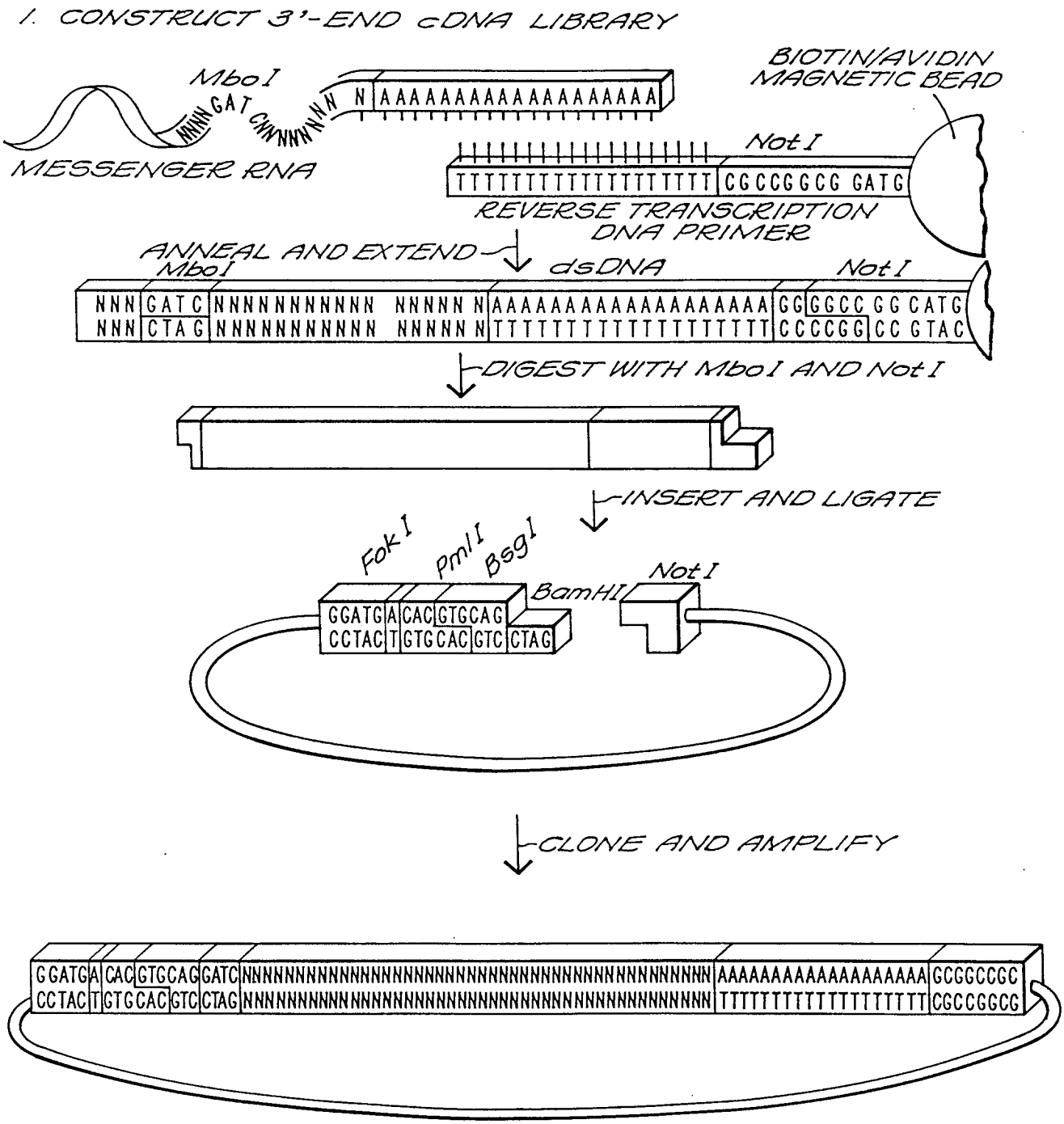

1. CONSTRUCT 3'-END cDNA LIBRARY
MboI
BIOTIN/AVIDIN MAGNETIC BEAD
MESSENGER RNA
NotI
REVERSE TRANSCRIPTION DNA PRIMER
ANNEAL AND EXTEND
dsDNA
DIGEST WITH MboI AND NotI
INSERT AND LIGATE
FokI
PmlI
BsgI
BamHI
NotI
CLONE AND AMPLIFY

FIG. 2A

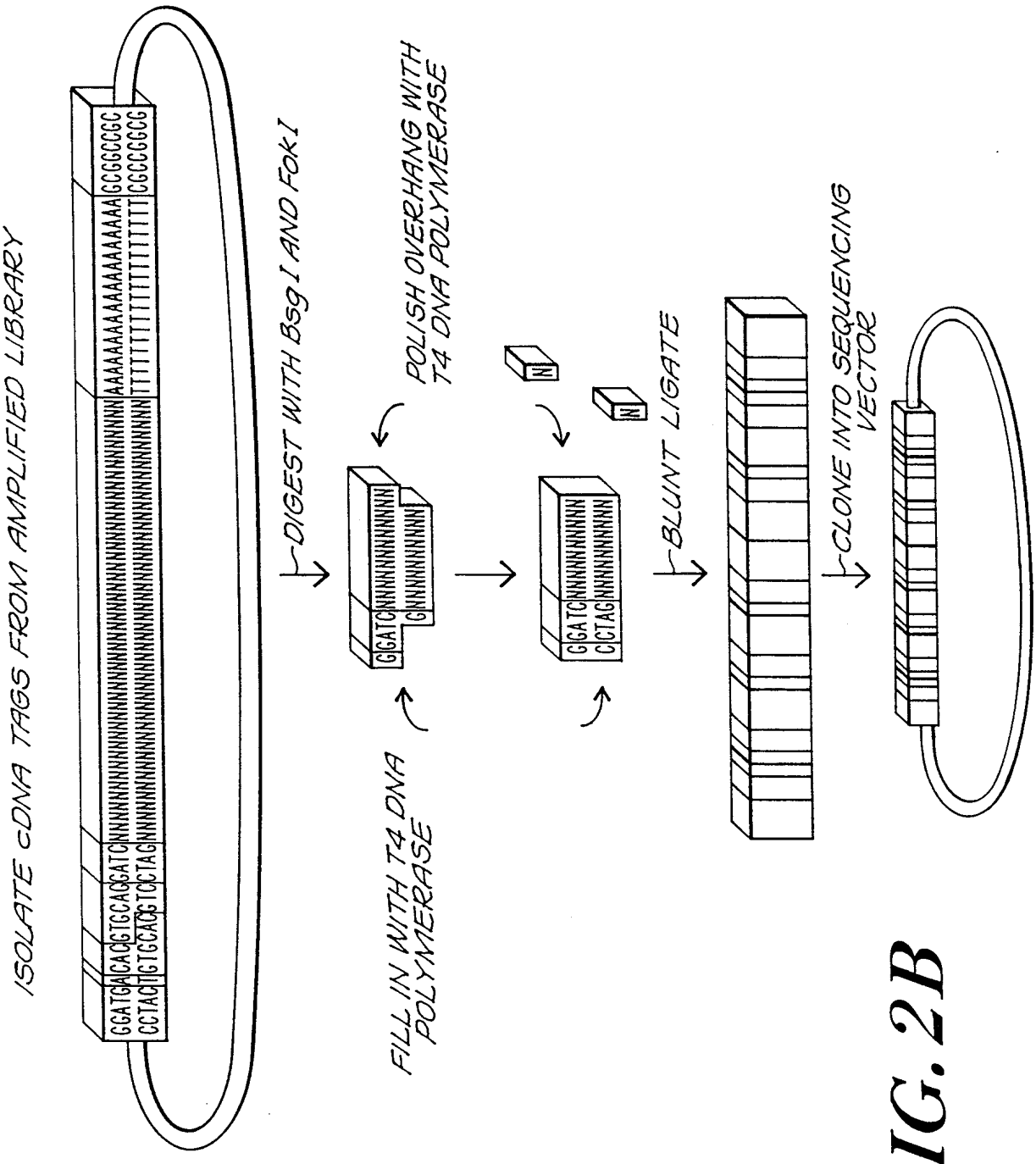

ISOLATE cDNA TAGS FROM AMPLIFIED LIBRARY
DIGEST WITH Bsg I AND FokI
POLISH OVERHANG WITH T4 DNA POLYMERASE
FILL IN WITH T4 DNA POLYMERASE
BLUNT LIGATE
CLONE INTO SEQUENCING VECTOR

FIG. 2B

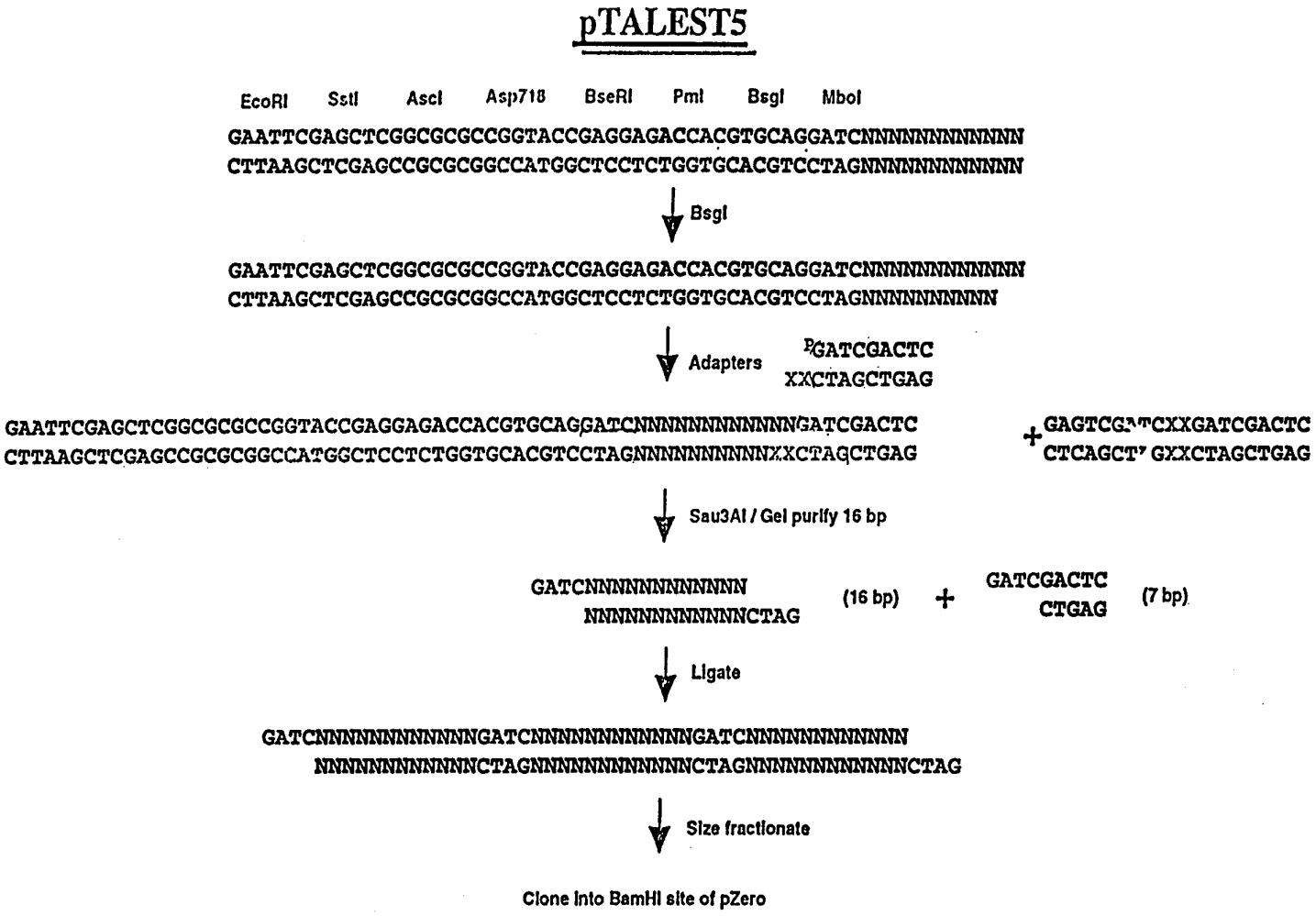

pTALEST5
EcoRI SstI AscI Asp718 BseRI PmI BsgI MboI
GAATTCGAGCTCGGCGCGCCGGTACCGAGGAGACCACGTGCAGGATCNNNNNNNNNNNNN
CTTAAGCTCGAGCCGCGCGGCCATGGCTCCTCTGGTGCACGTCCTAGNNNNNNNNNNNNN
BsgI
GAATTCGAGCTCGGCGCGCCGGTACCGAGGAGACCACGTGCAGGATCNNNNNNNNNNNNN
CTTAAGCTCGAGCCGCGCGGCCATGGCTCCTCTGGTGCACGTCCTAGNNNNNNNNNNNN
Adapters
GATCGACTC
XXCTAGCTGAG
GAATTCGAGCTCGGCGCGCCGGTACCGAGGAGACCACGTGCAGGATCNNNNNNNNNNNNNNGATCGACTC
CTTAAGCTCGAGCCGCGCGGCCATGGCTCCTCTGGTGCACGTCCTAGNNNNNNNNNNNNXXCTAGCTGAG
+ GAGTCGATCXXGATCGACTC
CTCAGCT GXXCTAGCTGAG
Sau3AI / Gel purify 16 bp
GATCNNNNNNNNNNNN
NNNNNNNNNNNNCTAG
(16 bp)
+
GATCGACTC
CTGAG
(7 bp)
Ligate
GATCNNNNNNNNNNNNNGATCNNNNNNNNNNNNNGATCNNNNNNNNNNNN
NNNNNNNNNNNNNCTAGNNNNNNNNNNNNNCTAGNNNNNNNNNNNNCTAG
Size fractionate
Clone into BamHI site of pZero

FIGURE 3

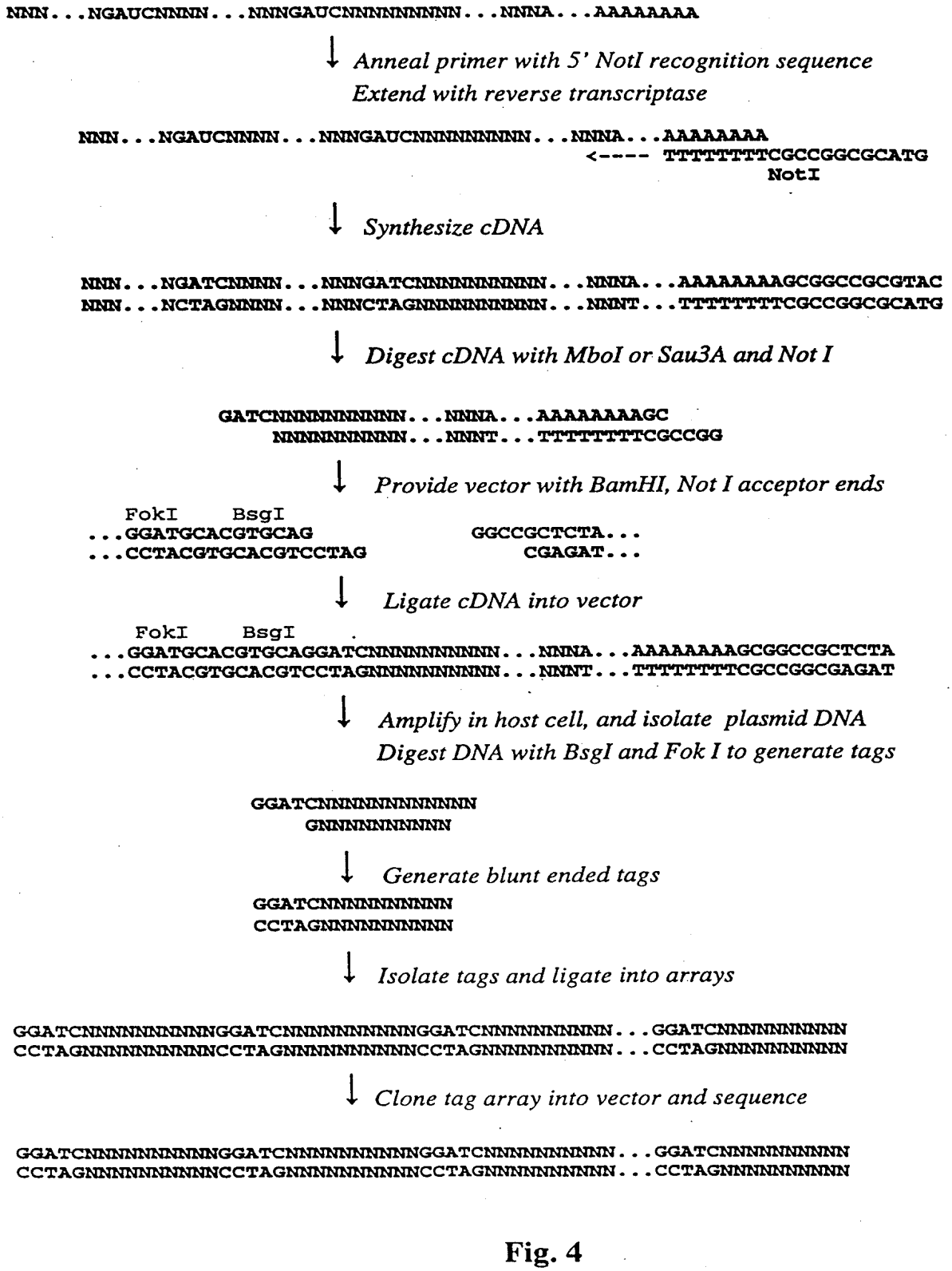

NNN...NGAUCNNNN...NNNGAUCNNNNNNNNN...NNNA...AAAAAAAA
Anneal primer with 5' NotI recognition sequence
Extend with reverse transcriptase
NNN...NGAUCNNNN...NNNGAUCNNNNNNNNN...NNNA...AAAAAAAA
<---- TTTTTTTTCGCCGGCGCATG
NotI
Synthesize cDNA
NNN...NGATCNNNN...NNNGATCNNNNNNNNNN...NNNA...AAAAAAAAGCGGCCGCGTAC
NNN...NCTAGNNNN...NNNCTAGNNNNNNNNNN...NNNT...TTTTTTTTCGCCGGCGCATG
Digest cDNA with MboI or Sau3A and Not I
GATCNNNNNNNNNN...NNNA...AAAAAAAAGC
NNNNNNNNNN...NNNT...TTTTTTTTCGCCGG
Provide vector with BamHI, Not I acceptor ends
FokI BsgI
...GGATGCACGTGCAG GGCCGCTCTA...
...CCTACGTGCACGTCCTAG CGAGAT...
Ligate cDNA into vector
FokI BsgI
...GGATGCACGTGCAGGATCNNNNNNNNNN...NNNA...AAAAAAAAGCGGCCGCTCTA
...CCTACGTGCACGTCCTAGNNNNNNNNNN...NNNT...TTTTTTTTCGCCGGCGAGAT
Amplify in host cell, and isolate plasmid DNA
Digest DNA with BsgI and Fok I to generate tags
GGATCNNNNNNNNNNNN
GNNNNNNNNNN
Generate blunt ended tags
GGATCNNNNNNNNNN
CCTAGNNNNNNNNNN
Isolate tags and ligate into arrays
GGATCNNNNNNNNNNGGATCNNNNNNNNNNGGATCNNNNNNNNNN...GGATCNNNNNNNNNN
CCTAGNNNNNNNNNNCCTAGNNNNNNNNNNCCTAGNNNNNNNNNN...CCTAGNNNNNNNNNN
Clone tag array into vector and sequence
GGATCNNNNNNNNNNGGATCNNNNNNNNNNGGATCNNNNNNNNNN...GGATCNNNNNNNNNN
CCTAGNNNNNNNNNNCCTAGNNNNNNNNNNCCTAGNNNNNNNNNN...CCTAGNNNNNNNNNN

Fig. 4

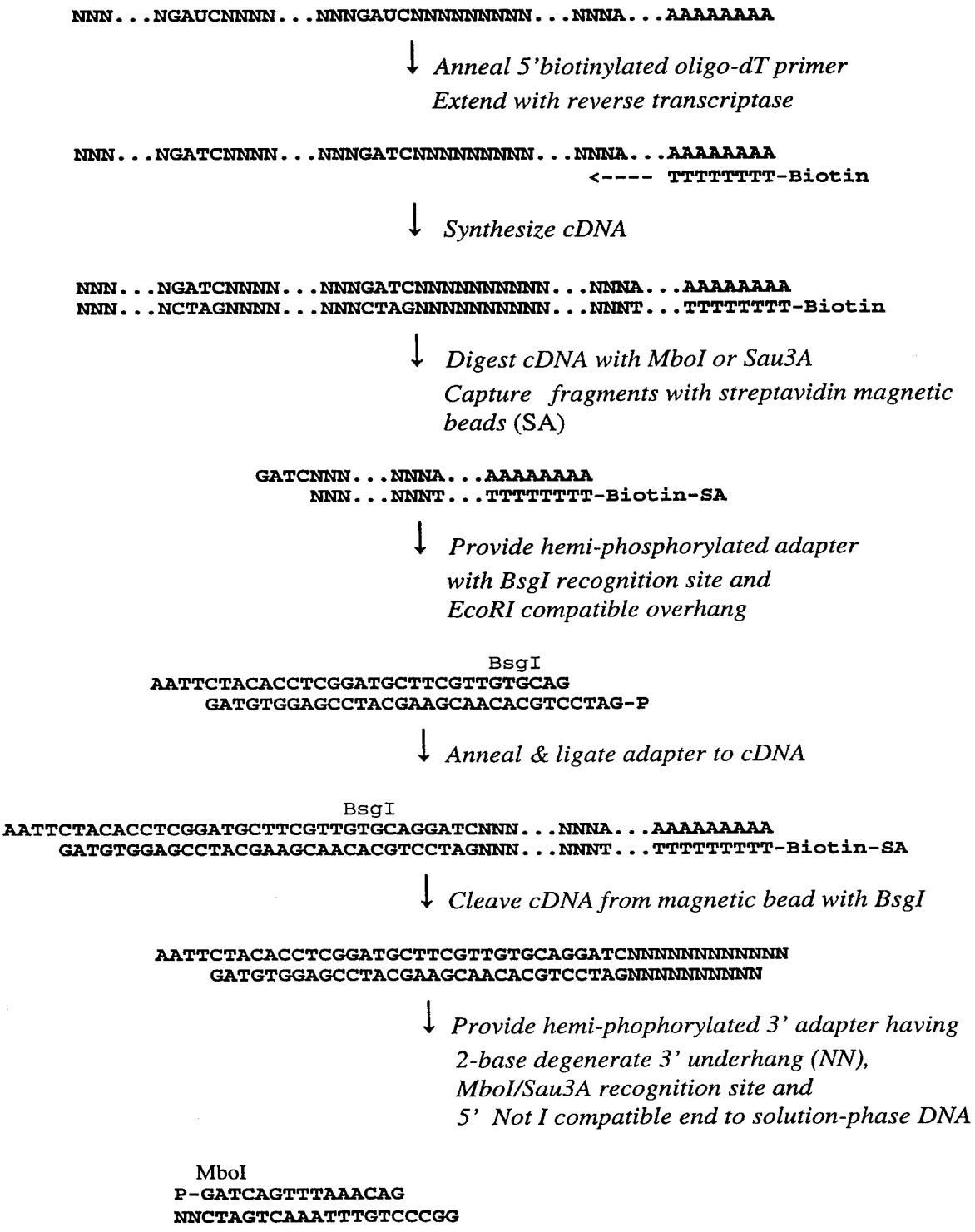

NNN...NGAUCNNNN...NNNGAUCNNNNNNNNN...NNNA...AAAAAAAA
Anneal 5'biotinylated oligo-dT primer
Extend with reverse transcriptase
NNN...NGATCNNNN...NNNGATCNNNNNNNNN...NNNA...AAAAAAAA
<---- TTTTTTTT-Biotin
Synthesize cDNA
NNN...NGATCNNNN...NNNGATCNNNNNNNNN...NNNA...AAAAAAAA
NNN...NCTAGNNNN...NNNCTAGNNNNNNNNN...NNNT...TTTTTTTT-Biotin
Digest cDNA with MboI or Sau3A
Capture fragments with streptavidin magnetic beads (SA)
GATCNNN...NNNA...AAAAAAAA
NNN...NNNT...TTTTTTTT-Biotin-SA
Provide hemi-phosphorylated adapter with BsgI recognition site and EcoRI compatible overhang
BsgI
AATTCTACACCTCGGATGCTTCGTTGTGCAG
GATGTGGAGCCTACGAAGCAACACGTCCTAG-P
Anneal & ligate adapter to cDNA
BsgI
AATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNN...NNNA...AAAAAAAAA
GATGTGGAGCCTACGAAGCAACACGTCCTAGNNN...NNNT...TTTTTTTTT-Biotin-SA
Cleave cDNA from magnetic bead with BsgI
AATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNNNNNNNNNNNN
GATGTGGAGCCTACGAAGCAACACGTCCTAGNNNNNNNNNNN
Provide hemi-phophorylated 3' adapter having 2-base degenerate 3' underhang (NN), MboI/Sau3A recognition site and 5' Not I compatible end to solution-phase DNA
MboI
P-GATCAGTTTAAACAG
NNCTAGTCAAATTTGTCCCGG

Fig. 5A

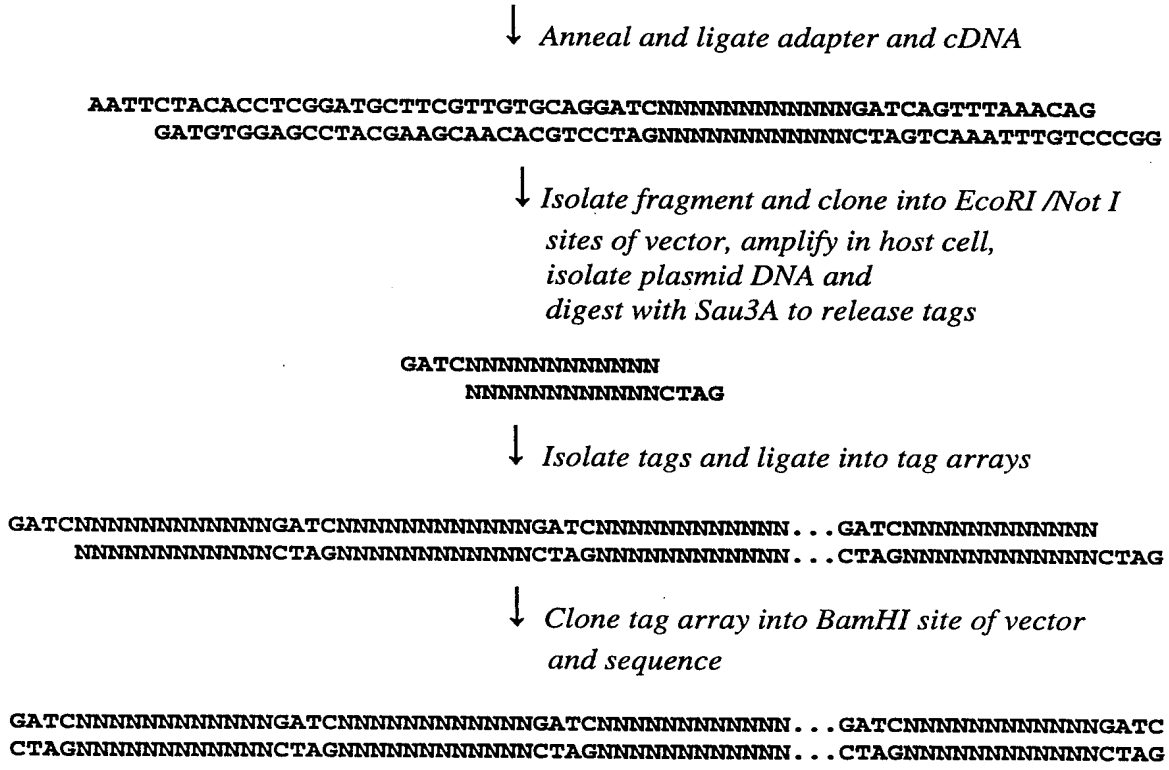

↓ Anneal and ligate adapter and cDNA
AATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNNNNNNNNNNNNGATCAGTTTAAACAG
GATGTGGAGCCTACGAAGCAACACGTCCTAGNNNNNNNNNNNNNNCTAGTCAAATTTGTCCCGG
↓ Isolate fragment and clone into EcoRI /Not I sites of vector, amplify in host cell, isolate plasmid DNA and digest with Sau3A to release tags
GATCNNNNNNNNNNNN
NNNNNNNNNNNNNCTAG
↓ Isolate tags and ligate into tag arrays
GATCNNNNNNNNNNNNNGATCNNNNNNNNNNNNNGATCNNNNNNNNNNNN...GATCNNNNNNNNNNNN
NNNNNNNNNNNNNCTAGNNNNNNNNNNNNNCTAGNNNNNNNNNNNN...CTAGNNNNNNNNNNNNCTAG
↓ Clone tag array into BamHI site of vector and sequence
GATCNNNNNNNNNNNNNGATCNNNNNNNNNNNNNGATCNNNNNNNNNNNN...GATCNNNNNNNNNNNNGATC
CTAGNNNNNNNNNNNNCTAGNNNNNNNNNNNNCTAGNNNNNNNNNNNN...CTAGNNNNNNNNNNNNCTAG

Fig. 5B

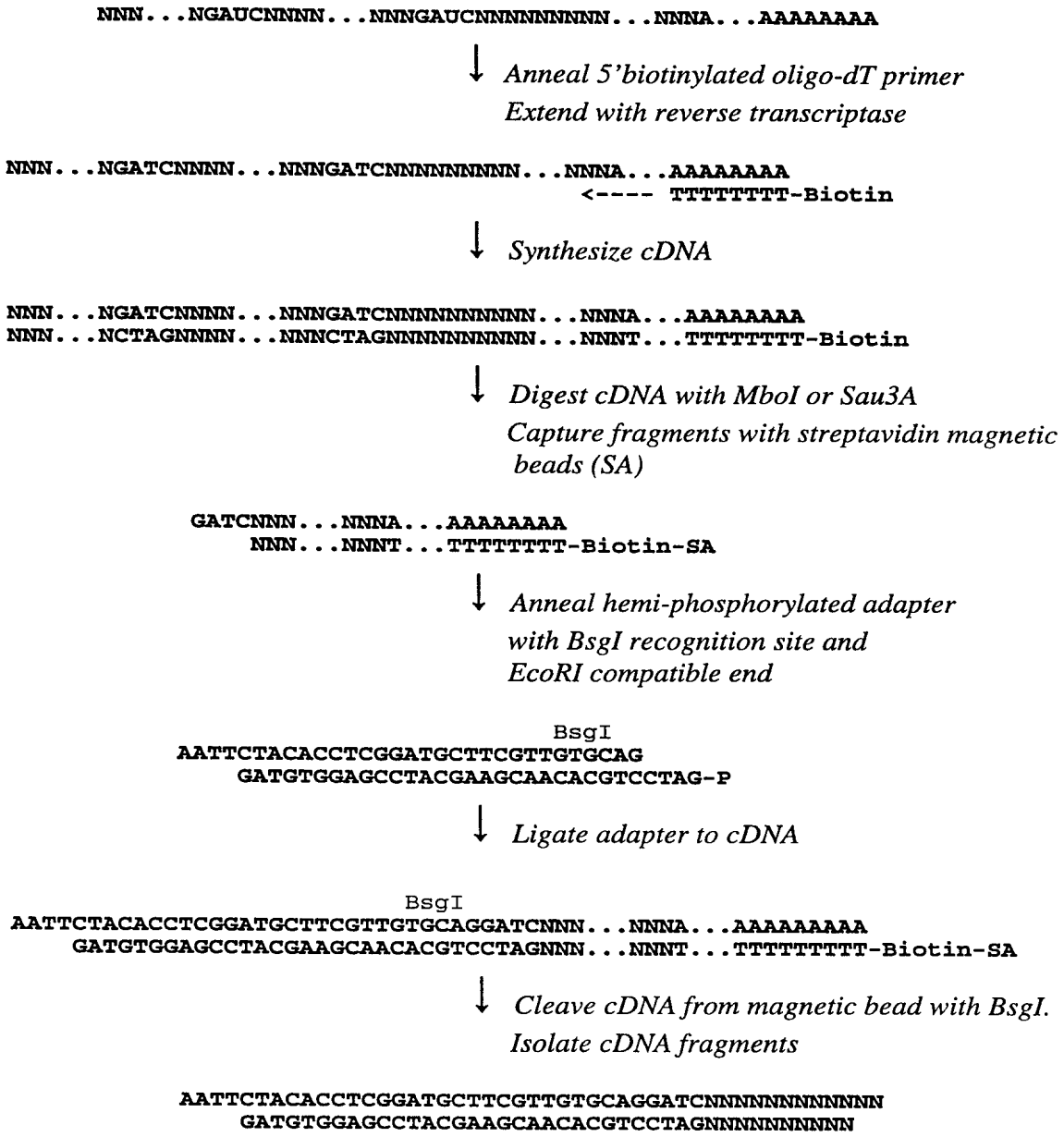

NNN...NGAUCNNNN...NNNGAUCNNNNNNNNNN...NNNA...AAAAAAAA
Anneal 5'biotinylated oligo-dT primer
Extend with reverse transcriptase
NNN...NGATCNNNN...NNNGATCNNNNNNNNNN...NNNA...AAAAAAAA
<---- TTTTTTTT-Biotin
Synthesize cDNA
NNN...NGATCNNNN...NNNGATCNNNNNNNNNN...NNNA...AAAAAAAA
NNN...NCTAGNNNN...NNNCTAGNNNNNNNNNN...NNNT...TTTTTTTT-Biotin
Digest cDNA with MboI or Sau3A
Capture fragments with streptavidin magnetic beads (SA)
GATCNNN...NNNA...AAAAAAAA
NNN...NNNT...TTTTTTTT-Biotin-SA
Anneal hemi-phosphorylated adapter with BsgI recognition site and EcoRI compatible end
BsgI
AATTCTACACCTCGGATGCTTCGTTGTGCAG
GATGTGGAGCCTACGAAGCAACACGTCCTAG-P
Ligate adapter to cDNA
BsgI
AATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNN...NNNA...AAAAAAAAA
GATGTGGAGCCTACGAAGCAACACGTCCTAGNNN...NNNT...TTTTTTTTT-Biotin-SA
Cleave cDNA from magnetic bead with BsgI.
Isolate cDNA fragments
AATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNNNNNNNNNNNN
GATGTGGAGCCTACGAAGCAACACGTCCTAGNNNNNNNNNNN

Fig. 6A

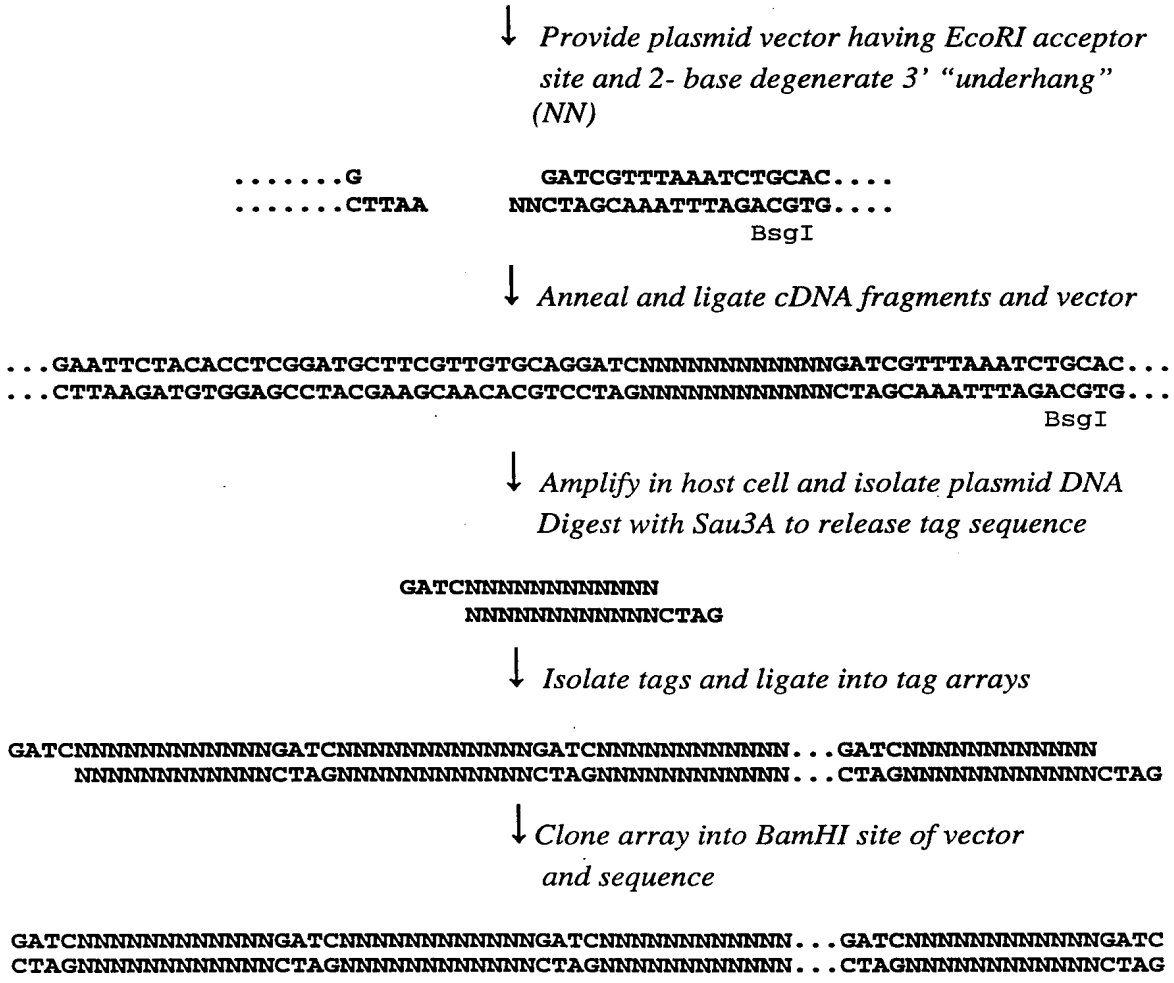

Provide plasmid vector having EcoRI acceptor site and 2- base degenerate 3' "underhang" (NN)
.....G GATCGTTTAAATCTGCAC....
.....CTTAA NNCTAGCAAATTTAGACGTG....
BsgI
Anneal and ligate cDNA fragments and vector
...GAATTCTACACCTCGGATGCTTCGTTGTGCAGGATCNNNNNNNNNNNNNGATCGTTTAAATCTGCAC...
...CTTAAGATGTGGAGCCTACGAAGCAACACGTCCTAGNNNNNNNNNNNNNCTAGCAAATTTAGACGTG...
BsgI
Amplify in host cell and isolate plasmid DNA
Digest with Sau3A to release tag sequence
GATCNNNNNNNNNNNN
NNNNNNNNNNNNCTAG
Isolate tags and ligate into tag arrays
GATCNNNNNNNNNNNNGATCNNNNNNNNNNNNGATCNNNNNNNNNNNN...GATCNNNNNNNNNNNN
NNNNNNNNNNNNCTAGNNNNNNNNNNNNCTAGNNNNNNNNNNNN...CTAGNNNNNNNNNNNNCTAG
Clone array into BamHI site of vector and sequence
GATCNNNNNNNNNNNNGATCNNNNNNNNNNNNGATCNNNNNNNNNNNN...GATCNNNNNNNNNNNNGATC
CTAGNNNNNNNNNNNNCTAGNNNNNNNNNNNNCTAGNNNNNNNNNNNN...CTAGNNNNNNNNNNNNCTAG

Fig. 6B

INTERNATIONAL SEARCH REPORT

Inte ional Application No
PCT/US 98/00965

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68 C12N15/10

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VELCULESCU ET AL.: "SERIAL ANALYSIS OF GENE EXPRESSION" SCIENCE, vol. 270, 1995, pages 484-487, XP002053721 cited in the application see the whole document ---	1-23
X	KATO: "DESCRIPTION OF THE ENTIRE mRNA POPULATION BY A 3'END cDNA FRAGMENT GENERATED BY CLASS IIS RESTRICTION ENZYMES" NUCLEIC ACID RESEARCH, vol. 23, no. 18, 1995, pages 3685-3690, XP002053720 see the whole document ---	1-23
	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 May 1998	08/06/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hagenmaier, S

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inte onal Application No
PCI/US 98/00965

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 508 169 A (DEUGAU KENNETH V ET AL) 16 April 1996 see the whole document ---	1-23
A	WO 95 20681 A (INCYTE PHARMA INC) 3 August 1995 see the whole document ---	1-23
A	UNRAU P ET AL: "NON-CLONING AMPLIFICATION OF SPECIFIC DNA FRAGMENTS FROM WHOLE GENOMIC DNA DIGEST USING DNA 'INDEXERS'" GENE, vol. 145, 1994, pages 163-169, XP002054436 see the whole document ---	1-23
A	ZAP EXPRESS cDNA SYNTHESIS KIT #200403 STRATAGENE 1995 XP002065732 see the whole document ---	1-23
P,X	EP 0 761 822 A (UNIV JOHNS HOPKINS MED) 12 March 1997 see the whole document ---	1-23
E	WO 98 14619 A (COCKS BENJAMIN GRAEME ;INCYTE PHARMA INC (US); CHUNG ALICIA (US);) 9 April 1998 see the whole document -----	1-23

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inte ional Application No
PCI/US 98/00965

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5508169 A	16-04-1996	CA 2036946 A	07-10-1991
WO 9520681 A	03-08-1995	AU 688465 B	12-03-1998
		AU 1694695 A	15-08-1995
		BG 100751 A	31-07-1997
		CA 2182217 A	03-08-1995
		CN 1145098 A	12-03-1997
		CZ 9602189 A	14-05-1997
		EP 0748390 A	18-12-1996
		FI 962987 A	26-09-1996
		JP 9503921 T	22-04-1997
		LV 11696 B	20-08-1997
		NO 963151 A	27-09-1996
		PL 315687 A	25-11-1996
		HU 75550 A	28-05-1997
EP 0761822 A	12-03-1997	US 5695937 A	09-12-1997
		AU 6561496 A	20-03-1997
		AU 7018896 A	01-04-1997
		CA 2185379 A	13-03-1997
		GB 2305241 A	02-04-1997
		WO 9710363 A	20-03-1997
WO 9814619 A	09-04-1998	NONE	